Cengage

Business Analytics

Descriptive • Predictive • Prescriptive



Camm Cochran Fry Ohlmann



Business Analytics

Descriptive • Predictive • Prescriptive

Jeffrey D. Camm
Wake Forest University

James J. Cochran
University of Alabama

Michael J. Fry
University of Cincinnati

Jeffrey W. Ohlmann
University of Iowa

Cengage

Australia • Brazil • Canada • Mexico • Singapore • United Kingdom • United States

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This is an electronic version of the print textbook. Due to electronic rights restrictions, some third party content may be suppressed. Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. The publisher reserves the right to remove content from this title at any time if subsequent rights restrictions require it. For valuable information on pricing, previous editions, changes to current editions, and alternate formats, please visit www.cengage.com/highered to search by ISBN#, author, title, or keyword for materials in your areas of interest.

Important Notice: Media content referenced within the product description or the product text may not be available in the eBook version.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Cengage

***Business Analytics,* Fifth Edition**
Jeffrey D. Camm, James J. Cochran, Michael J. Fry, Jeffrey W. Ohlmann

VP, Product Management & Marketing, Learning Experiences: Thais Alencarr

Product Director: Joe Sabatino

Senior Portfolio Product Manager: Aaron Arnsparger

Content Manager: Breanna Holmes

Product Assistant: Flannery Cowan

Marketing Manager: Colin Kramer

Senior Learning Designer: Brandon Foltz

Digital Product Manager: Andrew Southwell

Intellectual Property Analyst: Ashley Maynard

Intellectual Property Project Manager: Kelli Besse

Production Service: MPS Limited

Project Manager, MPS Limited: Shreya Tiwari

Art Director: Chris Doughman

Text Designer: Beckmeyer Design

Cover Designer: Beckmeyer Design

Cover Image: iStockPhoto.com/tawanlubfah

Copyright © 2024 Cengage Learning, Inc. ALL RIGHTS RESERVED.
WCN: 02-300
No part of this work covered by the copyright herein may be reproduced or distributed in any form or by any means, except as permitted by U.S. copyright law, without the prior written permission of the copyright owner.

Unless otherwise noted, all content is Copyright © Cengage Learning, Inc.

Previous Editions: ©2021, ©2019, ©2017

For product information and technology assistance, contact us at
Cengage Customer & Sales Support, 1-800-354-9706
or support.cengage.com.

For permission to use material from this text or product, submit all requests online at **www.copyright.com**.

Library of Congress Control Number: 2023900651

Student Edition:
ISBN: 978-0-357-90220-2

Loose-leaf Edition
ISBN: 978-0-357-90221-9

Cengage
200 Pier 4 Boulevard
Boston, MA 02210
USA

Cengage is a leading provider of customized learning solutions. Our employees reside in nearly 40 different countries and serve digital learners in 165 countries around the world. Find your local representative at: **www.cengage.com.**

To learn more about Cengage platforms and services, register or access your online learning solution, or purchase materials for your course, visit **www.cengage.com.**

Printed in the United States of America
Print Number: 01 Print Year: 2023

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Brief Contents

About the Authors xix
Preface xxi

Chapter 1	Introduction to Business Analytics 1
Chapter 2	Descriptive Statistics 21
Chapter 3	Data Visualization 83
Chapter 4	Data Wrangling: Data Management and Data Cleaning Strategies 151
Chapter 5	Probability: An Introduction to Modeling Uncertainty 199
Chapter 6	Descriptive Data Mining 257
Chapter 7	Statistical Inference 319
Chapter 8	Linear Regression 393
Chapter 9	Time Series Analysis and Forecasting 471
Chapter 10	Predictive Data Mining: Regression Tasks 523
Chapter 11	Predictive Data Mining: Classification Tasks 571
Chapter 12	Spreadsheet Models 633
Chapter 13	Monte Carlo Simulation 671
Chapter 14	Linear Optimization Models 743
Chapter 15	Integer Linear Optimization Models 805
Chapter 16	Nonlinear Optimization Models 849
Chapter 17	Decision Analysis 893
Multi-Chapter Case Problems	
	Capital State University Game-Day Magazines 943
	Hanover Inc. 945
Appendix A	Basics of Excel 947
Appendix B	Database Basics with Microsoft Access 959
Appendix C	Solutions to Even-Numbered Problems (Cengage eBook)
Appendix D	Microsoft Excel Online and Tools for Statistical Analysis (Cengage eBook)
References	*Available in the Cengage eBook*
Index	*997*

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Contents

About the Authors xix
Preface xxi

Chapter 1 Introduction to Business Analytics 1
1.1 Decision Making 3
1.2 Business Analytics Defined 3
1.3 A Categorization of Analytical Methods and Models 4
Descriptive Analytics 4
Predictive Analytics 5
Prescriptive Analytics 5
1.4 Big Data, the Cloud, and Artificial Intelligence 6
Volume 6
Velocity 6
Variety 7
Veracity 7
1.5 Business Analytics in Practice 9
Accounting Analytics 9
Financial Analytics 10
Human Resource (HR) Analytics 10
Marketing Analytics 10
Health Care Analytics 11
Supply Chain Analytics 11
Analytics for Government and Nonprofits 11
Sports Analytics 12
Web Analytics 12
1.6 Legal and Ethical Issues in the Use of Data and Analytics 12
Summary 15
Glossary 15
Problems 16
Available in the Cengage eBook:
Appendix: Getting Started with R and Rstudio
Appendix: Basic Data Manipulation with R

Chapter 2 Descriptive Statistics 21
2.1 Overview of Using Data: Definitions and Goals 23
2.2 Types of Data 24
Population and Sample Data 24
Quantitative and Categorical Data 24
Cross-Sectional and Time Series Data 24
Sources of Data 25
2.3 Exploring Data in Excel 27
Sorting and Filtering Data in Excel 27
Conditional Formatting of Data in Excel 30

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2.4 Creating Distributions from Data 32
Frequency Distributions for Categorical Data 32
Relative Frequency and Percent Frequency Distributions 33
Frequency Distributions for Quantitative Data 35
Histograms 37
Frequency Polygons 42
Cumulative Distributions 43
2.5 Measures of Location 46
Mean (Arithmetic Mean) 46
Median 47
Mode 48
Geometric Mean 48
2.6 Measures of Variability 51
Range 51
Variance 52
Standard Deviation 53
Coefficient of Variation 54
2.7 Analyzing Distributions 54
Percentiles 55
Quartiles 56
z-Scores 56
Empirical Rule 57
Identifying Outliers 59
Boxplots 59
2.8 Measures of Association Between Two Variables 62
Scatter Charts 62
Covariance 64
Correlation Coefficient 67
Summary 68
Glossary 69
Problems 70
Case Problem 1: Heavenly Chocolates Web Site Transactions 80
Case Problem 2: African Elephant Populations 81
Available in the Cengage eBook:
Appendix: Descriptive Statistics with R

Chapter 3 Data Visualization 83
3.1 Overview of Data Visualization 86
Preattentive Attributes 86
Data-Ink Ratio 89
3.2 Tables 92
Table Design Principles 93
Crosstabulation 94
PivotTables in Excel 97
3.3 Charts 101
Scatter Charts 101

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Recommended Charts in Excel 103
Line Charts 104
Bar Charts and Column Charts 108
A Note on Pie Charts and Three-Dimensional Charts 112
Additional Visualizations for Multiple Variables: Bubble Chart, Scatter Chart Matrix, and Table Lens 112
PivotCharts in Excel 117
3.4 Specialized Data Visualizations 120
Heat Maps 120
Treemaps 121
Waterfall Charts 122
Stock Charts 124
Parallel-Coordinates Chart 126
3.5 Visualizing Geospatial Data 126
Choropleth Maps 127
Cartograms 129
3.6 Data Dashboards 131
Principles of Effective Data Dashboards 131
Applications of Data Dashboards 132
Summary 134
Glossary 134
Problems 136
Case Problem 1: Pelican Stores 149
Case Problem 2: Movie Theater Releases 150
Available in the Cengage eBook:
Appendix: Creating Tabular and Graphical Presentations with R
Appendix: Data Visualization with Tableau

Chapter 4 Data Wrangling: Data Management and Data Cleaning Strategies 151
4.1 Discovery 153
Accessing Data 153
The Format of the Raw Data 157
4.2 Structuring 158
Data Formatting 159
Arrangement of Data 159
Splitting a Single Field into Multiple Fields 161
Combining Multiple Fields into a Single Field 165
4.3 Cleaning 167
Missing Data 167
Identification of Erroneous Outliers, Other Erroneous Values, and Duplicate Records 170
4.4 Enriching 176
Subsetting Data 177
Supplementing Data 179
Enhancing Data 182

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4.5 Validating and Publishing 186
Validating 186
Publishing 188
Summary 188
Glossary 189
Problems 190
Case Problem 1: Usman Solutions 197
Available in the Cengage eBook:
Appendix: Importing Delimited Files into R
Appendix: Working with Records in R
Appendix: Working with Fields in R
Appendix: Unstacking and Stacking Data in R

Chapter 5 Probability: An Introduction to Modeling Uncertainty 199
5.1 Events and Probabilities 201
5.2 Some Basic Relationships of Probability 202
Complement of an Event 202
Addition Law 203
5.3 Conditional Probability 205
Independent Events 210
Multiplication Law 210
Bayes' Theorem 211
5.4 Random Variables 213
Discrete Random Variables 213
Continuous Random Variables 214
5.5 Discrete Probability Distributions 215
Custom Discrete Probability Distribution 215
Expected Value and Variance 217
Discrete Uniform Probability Distribution 220
Binomial Probability Distribution 221
Poisson Probability Distribution 224
5.6 Continuous Probability Distributions 227
Uniform Probability Distribution 227
Triangular Probability Distribution 229
Normal Probability Distribution 231
Exponential Probability Distribution 236
Summary 240
Glossary 240
Problems 242
Case Problem 1: Hamilton County Judges 254
Case Problem 2: McNeil's Auto Mall 255
Case Problem 3: Gebhardt Electronics 256

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Available in the Cengage eBook:
Appendix: Discrete Probability Distributions with R
Appendix: Continuous Probability Distributions with R

Chapter 6 Descriptive Data Mining 257

6.1 Dimension Reduction 259
Geometric Interpretation of Principal Component Analysis 259
Summarizing Protein Consumption for Maillard Riposte 262

6.2 Cluster Analysis 266
Measuring Distance Between Observations Consisting of Quantitative Variables 267
Measuring Distance Between Observations Consisting of Categorical Variables 269
k-Means Clustering 271
Hierarchical Clustering and Measuring Dissimilarity Between Clusters 275
Hierarchical Clustering versus *k*-Means Clustering 283

6.3 Association Rules 284
Evaluating Association Rules 286

6.4 Text Mining 287
Voice of the Customer at Triad Airlines 288
Preprocessing Text Data for Analysis 289
Movie Reviews 290
Computing Dissimilarity Between Documents 293
Word Clouds 294

Summary 295
Glossary 296
Problems 298
Case Problem 1: Big Ten Expansion 315
Case Problem 2: Know Thy Customer 316
Available in the Cengage eBook:
Appendix: Principal Component Analysis with R
Appendix: *k*-Means Clustering with R
Appendix: Hierarchical Clustering with R
Appendix: Association Rules with R
Appendix: Text Mining with R
Appendix: Principal Component Analysis with Orange
Appendix: *k*-Means Clustering with Orange
Appendix: Hierarchical Clustering with Orange
Appendix: Association Rules with Orange
Appendix: Text Mining with Orange

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 7 Statistical Inference 319

7.1 Selecting a Sample 322
Sampling from a Finite Population 322
Sampling from an Infinite Population 323

7.2 Point Estimation 326
Practical Advice 328

7.3 Sampling Distributions 328
Sampling Distribution of $\bar{x}$ 331
Sampling Distribution of $\bar{p}$ 336

7.4 Interval Estimation 339
Interval Estimation of the Population Mean 339
Interval Estimation of the Population Proportion 346

7.5 Hypothesis Tests 349
Developing Null and Alternative Hypotheses 349
Type I and Type II Errors 352
Hypothesis Test of the Population Mean 353
Hypothesis Test of the Population Proportion 364

7.6 Big Data, Statistical Inference, and Practical Significance 367
Sampling Error 367
Nonsampling Error 368
Big Data 369
Understanding What Big Data Is 370
Big Data and Sampling Error 371
Big Data and the Precision of Confidence Intervals 372
Implications of Big Data for Confidence Intervals 373
Big Data, Hypothesis Testing, and *p* Values 374
Implications of Big Data in Hypothesis Testing 376

Summary 376
Glossary 377
Problems 380
Case Problem 1: Young Professional Magazine 390
Case Problem 2: Quality Associates, Inc. 391
Available in the Cengage eBook:
Appendix: Random Sampling with R
Appendix: Interval Estimation with R
Appendix: Hypothesis Testing with R

Chapter 8 Linear Regression 393

8.1 Simple Linear Regression Model 395
Estimated Simple Linear Regression Equation 395

8.2 Least Squares Method 397
Least Squares Estimates of the Simple Linear Regression Parameters 399

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Using Excel's Chart Tools to Compute the Estimated Simple Linear Regression Equation 401
8.3 Assessing the Fit of the Simple Linear Regression Model 403
The Sums of Squares 403
The Coefficient of Determination 405
Using Excel's Chart Tools to Compute the Coefficient of Determination 406
8.4 The Multiple Linear Regression Model 407
Estimated Multiple Linear Regression Equation 407
Least Squares Method and Multiple Linear Regression 408
Butler Trucking Company and Multiple Linear Regression 408
Using Excel's Regression Tool to Develop the Estimated Multiple Linear Regression Equation 409
8.5 Inference and Linear Regression 412
Conditions Necessary for Valid Inference in the Least Squares Linear Regression Model 413
Testing Individual Linear Regression Parameters 417
Addressing Nonsignificant Independent Variables 420
Multicollinearity 421
8.6 Categorical Independent Variables 424
Butler Trucking Company and Rush Hour 424
Interpreting the Parameters 426
More Complex Categorical Variables 427
8.7 Modeling Nonlinear Relationships 429
Quadratic Regression Models 430
Piecewise Linear Regression Models 434
Interaction Between Independent Variables 436
8.8 Model Fitting 441
Variable Selection Procedures 441
Overfitting 442
8.9 Big Data and Linear Regression 443
Inference and Very Large Samples 443
Model Selection 446
8.10 Prediction with Linear Regression 447
Summary 450
Glossary 450
Problems 452
Case Problem 1: Alumni Giving 466
Case Problem 2: Consumer Research, Inc. 468
Case Problem 3: Predicting Winnings for NASCAR Drivers 469
Available in the Cengage eBook:
Appendix: Simple Linear Regression with R
Appendix: Multiple Linear Regression with R
Appendix: Linear Regression Variable Selection Procedures with R

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 9 Time Series Analysis and Forecasting 471
9.1 Time Series Patterns 474
Horizontal Pattern 474
Trend Pattern 476
Seasonal Pattern 477
Trend and Seasonal Pattern 478
Cyclical Pattern 481
Identifying Time Series Patterns 481
9.2 Forecast Accuracy 481
9.3 Moving Averages and Exponential Smoothing 485
Moving Averages 486
Exponential Smoothing 490
9.4 Using Linear Regression Analysis for Forecasting 494
Linear Trend Projection 494
Seasonality Without Trend 496
Seasonality with Trend 497
Using Linear Regression Analysis as a Causal Forecasting Method 500
Combining Causal Variables with Trend and Seasonality Effects 503
Considerations in Using Linear Regression in Forecasting 504
9.5 Determining the Best Forecasting Model to Use 504
Summary 505
Glossary 505
Problems 506
Case Problem 1: Forecasting Food and Beverage Sales 515
Case Problem 2: Forecasting Lost Sales 515
Appendix 9.1: Using the Excel Forecast Sheet 517
Available in the Cengage eBook:
Appendix: Forecasting with R

Chapter 10 Predictive Data Mining: Regression Tasks 523
10.1 Regression Performance Measures 524
10.2 Data Sampling, Preparation, and Partitioning 526
Static Holdout Method 526
k-Fold Cross-Validation 530
10.3 *k*-Nearest Neighbors Regression 535
10.4 Regression Trees 538
Constructing a Regression Tree 538
Generating Predictions with a Regression Tree 541
Ensemble Methods 543
10.5 Neural Network Regression 548
Structure of a Neural Network 548
How a Neural Network Learns 552

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

10.6 Feature Selection 555
Wrapper Methods 556
Filter Methods 556
Embedded Methods 557
Summary 558
Glossary 558
Problems 560
Case Problem: Housing Bubble 568
Available in the Cengage eBook:
Appendix: *k*-Nearest Neighbors Regression with R
Appendix: Individual Regression Trees with R
Appendix: Random Forests of Regression Trees with R
Appendix: Neural Network Regression with R
Appendix: Regularized Linear Regression with R
Appendix: *k*-Nearest Neighbors Regression with Orange
Appendix: Individual Regression Trees with Orange
Appendix: Random Forests of Regression Trees with Orange
Appendix: Neural Network Regression with Orange
Appendix: Regularized Linear Regression with Orange

Chapter 11 Predictive Data Mining: Classification Tasks 571
11.1 Data Sampling, Preparation, and Partitioning 573
Static Holdout Method 573
k-Fold Cross-Validation 574
Class Imbalanced Data 574
11.2 Performance Measures for Binary Classification 576
11.3 Classification with Logistic Regression 582
11.4 *k*-Nearest Neighbors Classification 587
11.5 Classification Trees 591
Constructing a Classification Tree 591
Generating Predictions with a Classification Tree 593
Ensemble Methods 594
11.6 Neural Network Classification 600
Structure of a Neural Network 601
How a Neural Network Learns 605
11.7 Feature Selection 609
Wrapper Methods 609
Filter Methods 610
Embedded Methods 610
Summary 612
Glossary 612
Problems 615
Case Problem: Grey Code Corporation 630

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Available in the Cengage eBook:
Appendix: Classification via Logistic Regression with R
Appendix: *k*-Nearest Neighbors Classification with R
Appendix: Individual Classification Trees with R
Appendix: Random Forests of Classification Trees with R
Appendix: Neural Network Classification with R
Appendix: Classification via Logistic Regression with Orange
Appendix: *k*-Nearest Neighbors Classification with Orange
Appendix: Individual Classification Trees with Orange
Appendix: Random Forests of Classification Trees with Orange
Appendix: Neural Network Classification with Orange

Chapter 12 Spreadsheet Models 633
12.1 Building Good Spreadsheet Models 635
Influence Diagrams 635
Building a Mathematical Model 635
Spreadsheet Design and Implementing the Model in a Spreadsheet 637
12.2 What-If Analysis 640
Data Tables 640
Goal Seek 642
Scenario Manager 644
12.3 Some Useful Excel Functions for Modeling 649
SUM and SUMPRODUCT 650
IF and COUNTIF 651
XLOOKUP 654
12.4 Auditing Spreadsheet Models 656
Trace Precedents and Dependents 656
Show Formulas 656
Evaluate Formulas 658
Error Checking 658
Watch Window 659
12.5 Predictive and Prescriptive Spreadsheet Models 660
Summary 661
Glossary 661
Problems 662
Case Problem: Retirement Plan 670

Chapter 13 Monte Carlo Simulation 671
13.1 Risk Analysis for Sanotronics LLC 673
Base-Case Scenario 673
Worst-Case Scenario 674
Best-Case Scenario 674
Sanotronics Spreadsheet Model 674

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Use of Probability Distributions to Represent Random Variables 676
Generating Values for Random Variables with Excel 677
Executing Simulation Trials with Excel 681
Measuring and Analyzing Simulation Output 682
13.2 Inventory Policy Analysis for Promus Corp 686
Spreadsheet Model for Promus 687
Generating Values for Promus Corp's Demand 688
Executing Simulation Trials and Analyzing Output 691
13.3 Simulation Modeling for Land Shark Inc. 693
Spreadsheet Model for Land Shark 694
Generating Values for Land Shark's Random Variables 696
Executing Simulation Trials and Analyzing Output 698
Generating Bid Amounts with Fitted Distributions 700
13.4 Simulation with Dependent Random Variables 709
Spreadsheet Model for Press Teag Worldwide 709
13.5 Simulation Considerations 714
Verification and Validation 714
Advantages and Disadvantages of Using Simulation 714
Summary 715
Summary of Steps for Conducting a Simulation Analysis 715
Glossary 716
Problems 717
Case Problem 1: Four Corners 731
Case Problem 2: Ginsberg's Jewelry Snowfall Promotion 732
Appendix 13.1: Common Probability Distributions for Simulation 734

Chapter 14 Linear Optimization Models 743
14.1 A Simple Maximization Problem 745
Problem Formulation 746
Mathematical Model for the Par, Inc. Problem 748
14.2 Solving the Par, Inc. Problem 749
The Geometry of the Par, Inc. Problem 749
Solving Linear Programs with Excel Solver 751
14.3 A Simple Minimization Problem 755
Problem Formulation 755
Solution for the M&D Chemicals Problem 755
14.4 Special Cases of Linear Program Outcomes 757
Alternative Optimal Solutions 758
Infeasibility 759
Unbounded 760
14.5 Sensitivity Analysis 762
Interpreting Excel Solver Sensitivity Report 762
14.6 General Linear Programming Notation and More Examples 764
Investment Portfolio Selection 765

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Transportation Planning 768
Maximizing Banner Ad Revenue 772
Assigning Project Leaders to Clients 776
Diet Planning 779
14.7 Generating an Alternative Optimal Solution for a Linear Program 782
Summary 783
Glossary 784
Problems 785
Case Problem1: Investment Strategy 801
Case Problem 2: Solutions Plus 802
Available in the Cengage eBook:
Appendix: Linear Programming with R

Chapter 15 Integer Linear Optimization Models 805
15.1 Types of Integer Linear Optimization Models 806
15.2 Eastborne Realty, an Example of Integer Optimization 807
The Geometry of Linear All-Integer Optimization 808
15.3 Solving Integer Optimization Problems with Excel Solver 810
A Cautionary Note About Sensitivity Analysis 813
15.4 Applications Involving Binary Variables 815
Capital Budgeting 815
Fixed Cost 816
Bank Location 820
Product Design and Market Share Optimization 822
15.5 Modeling Flexibility Provided by Binary Variables 825
Multiple-Choice and Mutually Exclusive Constraints 825
k Out of *n* Alternatives Constraint 826
Conditional and Corequisite Constraints 826
15.6 Generating Alternatives in Binary Optimization 827
Summary 829
Glossary 830
Problems 830
Case Problem 1: Applecore Children's Clothing 845
Case Problem 2: Yeager National Bank 847
Available in the Cengage eBook:
Appendix: Integer Programming with R

Chapter 16 Nonlinear Optimization Models 849
16.1 A Production Application: Par, Inc. Revisited 850
An Unconstrained Problem 850
A Constrained Problem 851

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Solving Nonlinear Optimization Models Using Excel Solver 853
Sensitivity Analysis and Shadow Prices in Nonlinear Models 855
16.2 Local and Global Optima 856
Overcoming Local Optima with Excel Solver 858
16.3 A Location Problem 860
16.4 Markowitz Portfolio Model 861
16.5 Adoption of a New Product: The Bass Forecasting Model 866
16.6 Heuristic Optimization Using Excel's Evolutionary Method 869
Summary 877
Glossary 877
Problems 878
Case Problem: Portfolio Optimization with Transaction Costs 889
Available in the Cengage eBook:
Appendix: Nonlinear Programming with R

Chapter 17 Decision Analysis 893
17.1 Problem Formulation 895
Payoff Tables 896
Decision Trees 896
17.2 Decision Analysis Without Probabilities 897
Optimistic Approach 897
Conservative Approach 898
Minimax Regret Approach 898
17.3 Decision Analysis with Probabilities 900
Expected Value Approach 900
Risk Analysis 902
Sensitivity Analysis 903
17.4 Decision Analysis with Sample Information 904
Expected Value of Sample Information 909
Expected Value of Perfect Information 909
17.5 Computing Branch Probabilities with Bayes' Theorem 910
17.6 Utility Theory 913
Utility and Decision Analysis 914
Utility Functions 918
Exponential Utility Function 921
Summary 923
Glossary 923
Problems 925

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Case Problem 1: Property Purchase Strategy 939
Case Problem 2: Semiconductor Fabrication at Axeon Labs 941

Multi-Chapter Case Problems
Capital State University Game-Day Magazines 943
Hanover Inc. 945

Appendix A Basics of Excel 947

Appendix B Database Basics with Microsoft Access 959

Appendix C Solutions to Even-Numbered Problems (Cengage eBook)

Appendix D Microsoft Excel Online and Tools for Statistical Analysis (Cengage eBook)

References Available in the Cengage eBook

Index 997

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

About the Authors

Jeffrey D. Camm. Jeffrey D. Camm is the Inmar Presidential Chair in Analytics and Senior Associate Dean of Faculty in the School of Business at Wake Forest University. Born in Cincinnati, Ohio, he holds a B.S. from Xavier University (Ohio) and a Ph.D. from Clemson University. Prior to joining the faculty at Wake Forest, he was on the faculty of the University of Cincinnati. He has also been a visiting scholar at Stanford University and a visiting professor of business administration at the Tuck School of Business at Dartmouth College.

Dr. Camm has published over 45 papers in the general area of optimization applied to problems in operations management and marketing. He has published his research in *Science, Management Science, Operations Research, Interfaces*, and other professional journals. Dr. Camm was named the Dornoff Fellow of Teaching Excellence at the University of Cincinnati and he was the recipient of the 2006 INFORMS Prize for the Teaching of Operations Research Practice. A firm believer in practicing what he preaches, he has served as an operations research consultant to numerous companies and government agencies. From 2005 to 2010, he served as editor-in-chief of *INFORMS Journal on Applied Analytics* (formerly *Interfaces*). In 2017, he was named an INFORMS Fellow.

James J. Cochran. James J. Cochran is Professor of Applied Statistics, the Rogers-Spivey Faculty Fellow, and Associate Dean for Faculty and Research at the University of Alabama. Born in Dayton, Ohio, he earned his B.S., M.S., and M.B.A. degrees from Wright State University and his Ph.D. from the University of Cincinnati. He has been at the University of Alabama since 2014 and has been a visiting scholar at Stanford University, Universidad de Talca, the University of South Africa, and Pole Universitaire Leonard de Vinci.

Professor Cochran has published over 50 papers in the development and application of operations research and statistical methods. He has published his research in *Management Science, The American Statistician, Communications in Statistics—Theory and Methods, Annals of Operations Research, European Journal of Operational Research, Journal of Combinatorial Optimization, INFORMS Journal of Applied Analytics, Statistics and Probability Letters*, and other professional journals. He was the 2008 recipient of the INFORMS Prize for the Teaching of Operations Research Practice and the 2010 recipient of the Mu Sigma Rho Statistical Education Award. Professor Cochran was elected to the International Statistics Institute in 2005 and named a Fellow of the American Statistical Association in 2011. He received the Founders Award in 2014 and the Karl E. Peace Award in 2015 from the American Statistical Association. In 2017 he received the American Statistical Association's Waller Distinguished Teaching Career Award and was named a Fellow of INFORMS, and in 2018 he received the INFORMS President's Award. He has twice been recognized as a finalist for the Innovative Applications in Analytics Award.

A strong advocate for effective statistics and operations research education as a means of improving the quality of applications to real problems, Professor Cochran has organized and chaired teaching effectiveness workshops in Montevideo, Uruguay; Cape Town, South Africa; Cartagena, Colombia; Jaipur, India; Buenos Aires, Argentina; Nairobi, Kenya; Buea, Cameroon; Kathmandu, Nepal; Osijek, Croatia; Havana, Cuba; Ulaanbaatar, Mongolia; Chişinău, Moldova; Dar es Salaam, Tanzania; Sozopol, Bulgaria; Tunis, Tunisia; and Saint George's, Grenada. He has served as an operations research consultant to numerous companies and not-for-profit organizations. He served as editor-in-chief of *INFORMS Transactions on Education* from 2006 to 2012 and is on the editorial board of *INFORMS Journal on Applied Analytics* (formerly *Interfaces*), *International Transactions in Operational Research,* and *Significance*.

Michael J. Fry. Michael J. Fry is Professor of Operations, Business Analytics, and Information Systems and Academic Director of the Center for Business Analytics in the Carl H. Lindner College of Business at the University of Cincinnati. Born in Killeen, Texas, he earned a B.S. from Texas A&M University and M.S.E. and Ph.D. degrees from the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

University of Michigan. He has been at the University of Cincinnati since 2002, where he was previously Department Head and has been named a Lindner Research Fellow. He has also been a visiting professor at the Samuel Curtis Johnson Graduate School of Management at Cornell University and the Sauder School of Business at the University of British Columbia.

Professor Fry has published more than 25 research papers in journals such as *Operations Research, M&SOM, Transportation Science, Naval Research Logistics, IISE Transactions, Critical Care Medicine,* and *INFORMS Journal on Applied Analytics* (formerly *Interfaces*). His research has been funded by the National Science Foundation and other funding agencies. His research interests are in applying quantitative management methods to the areas of supply chain analytics, sports analytics, and public-policy operations. He has worked with many different organizations for his research, including Dell, Inc., Starbucks Coffee Company, Great American Insurance Group, the Cincinnati Fire Department, the State of Ohio Election Commission, the Cincinnati Bengals, and the Cincinnati Zoo & Botanical Garden. He was named a finalist for the Daniel H. Wagner Prize for Excellence in Operations Research Practice, and he has been recognized for both his research and teaching excellence at the University of Cincinnati.

Jeffrey W. Ohlmann. Jeffrey W. Ohlmann is Associate Professor of Management Sciences and Huneke Research Fellow in the Tippie College of Business at the University of Iowa. Born in Valentine, Nebraska, he earned a B.S. from the University of Nebraska, and M.S. and Ph.D. degrees from the University of Michigan. He has been at the University of Iowa since 2003.

Professor Ohlmann's research on the modeling and solution of decision-making problems has produced more than two dozen research papers in journals such as *Operations Research, Mathematics of Operations Research, INFORMS Journal on Computing, Transportation Science,* the *European Journal of Operational Research,* and *INFORMS Journal on Applied Analytics* (formerly *Interfaces*). He has collaborated with companies such as Transfreight, LeanCor, Cargill, the Hamilton County Board of Elections, and three National Football League franchises. Because of the relevance of his work to industry, he was bestowed the George B. Dantzig Dissertation Award and was recognized as a finalist for the Daniel H. Wagner Prize for Excellence in Operations Research Practice.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Preface

Business Analytics 5E is designed to introduce the concept of business analytics to undergraduate and graduate students. This edition builds upon what was one of the first collections of materials that are essential to the growing field of business analytics. In Chapter 1, we present an overview of business analytics and our approach to the material in this textbook. In simple terms, business analytics helps business professionals make better decisions based on data. We discuss exploring, wrangling, summarizing, visualizing, and understanding data in Chapters 2 through 5. In Chapter 6, we introduce additional models to describe data and the relationships among variables. Chapters 7 through 11 introduce methods for both gaining insights from historical data and predicting possible future outcomes. Chapter 12 covers the use of spreadsheets for examining data and building decision models. In Chapter 13, we demonstrate how to explicitly introduce uncertainty into spreadsheet models through the use of Monte Carlo simulation. In Chapters 14 through 16, we discuss optimization models to help decision makers choose the best decision based on the available data. Chapter 17 is an overview of decision analysis approaches for incorporating a decision maker's views about risk into decision making. In Appendix A we present optional material for students who need to learn the basics of using Microsoft Excel. The use of databases and manipulating data in Microsoft Access is discussed in Appendix B. Appendixes in many chapters illustrate the use of additional software tools such as Tableau, R, and Orange to apply analytics methods.

This textbook can be used by students who have previously taken a course on basic statistical methods as well as students who have not had a prior course in statistics. *Business Analytics* 5E is also amenable to a two-course sequence in business statistics and analytics. All statistical concepts contained in this textbook are presented from a business analytics perspective using practical business examples. Chapters 2, 5, 7, 8, and 9 provide an introduction to basic statistical concepts that form the foundation for more advanced analytics methods. Chapter 2 introduces descriptive statistical measures such as measures of location, measures of variability, methods to analyze distributions of data, and measures of association between variables. Chapter 5 covers material related to probability including discussions of discrete probability distributions and continuous probability distributions. Chapter 7 contains topics related to statistical inference including interval estimation and hypothesis testing. Chapter 8 explores linear regression models including simple linear models and multiple linear regression models, and chapter 9 covers methods for forecasting and time series data. Chapter 3 covers additional data visualization topics and Chapter 4 is an overview of approaches for exploring, cleaning, and wrangling data to make the data more amenable for analysis. These topics are not always covered in traditional business statistics courses, but they are essential in understanding how to analyze data. Chapters 6, 10, and 11 cover additional topics in data mining that are not traditionally part of most introductory business statistics courses, but they are exceedingly important and commonly used in current business environments. Chapter 6 focuses on data mining methods used to describe data and the relationships among variables such as cluster analysis, association rules and text mining. Chapter 6 also contains an overview of dimension reduction techniques such as principal component analysis. Chapters 10 and 11 focus on data mining methods used to make predictions from data. Chapter 10 covers models for regression tasks such as *k*-nearest neighbors, regression trees, and neural networks. Chapter 11 covers models for classification tasks such as logistic regression, *k*-nearest neighbors, classification trees, ensemble methods, and neural networks. Both chapters 10 and 11 cover data sampling methods for prediction models including *k*-fold cross-validation, and also provide an overview of feature selection methods such as wrapper methods, filter methods, and embedded methods (including regularization). Chapter 12 and Appendix A provide the foundational knowledge students need to use Microsoft Excel for analytics applications. Chapters 13 through 17 build upon this spreadsheet knowledge to present additional topics that are used by many organizations that use prescriptive analytics to improve decision making.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Updates in the Fifth Edition

The fifth edition of *Business Analytics* is a major revision that introduces new chapters, new concepts, and new tools. Chapter 4 is a new chapter on data wrangling that covers topics such as how to access and structure data for exploration, how to clean and enrich data to facilitate analysis, and how to validate data. Our coverage of data mining topics has been greatly expanded. Our coverage of descriptive data mining techniques in Chapter 6 now includes a discussion of how to conduct dimension reduction with principal component analysis (PCA), and we have thoroughly updated our coverage of clustering, association rules, and text mining. Our coverage of predictive data mining techniques now includes two separate chapters: Chapter 10 focuses on predicting quantitative outcomes with *k*-nearest neighbors regression, regression trees, ensemble methods, and neural network regression, while Chapter 11 focuses on predicting binary categorical outcomes with *k*-nearest neighbors classification, classification trees, ensemble methods, and neural network classification. In addition, we now include online appendixes that introduce the software package Orange for descriptive and predictive data-mining models. Orange is an open-source machine learning and data visualization software package built using Python. This coverage of Orange and Python complements our existing coverage of R for descriptive and predictive analytics. Our coverage of descriptive analytics methods in Chapter 2 and data visualization in Chapter 3 has been greatly expanded to provide additional depth and introduce new methods related to these topics. We have also increased the size of many data sets in Chapter 8 on linear regression and Chapter 9 on time series analysis and forecasting to better represent realistic data sets that are encountered in practice. We have added additional topics to our coverage of optimization models including a new section in Chapter 16 on heuristic optimization with the Excel Solver Evolutionary method. We also now provide a new online Appendix D that covers the basics of the use of Microsoft Excel Online for statistical analysis. Finally, we have added learning objectives (LOs) to the beginning of each chapter that explain the key concepts covered in each chapter, and we map these LOs to the end-of-chapter problems and cases.

- **New Chapter on Data Wrangling.** Chapter 4 covers methods for accessing and structuring data for exploring, cleaning and enriching data to facilitate analysis, and validating data. Many professionals in analytics and data science spend much of their time preparing data for exploration and analysis. It is essential that students are familiar with these methods. We have also created new online appendixes covering how to implement these methods in R.
- **New Material in Descriptive Data Mining Chapter.** Chapter 6 on descriptive data mining techniques now provides a more in-depth discussion of dimension reduction techniques such as principal component analysis. We present material to help students understand what this approach is doing, as well as how to interpret the output from this method. We have also rewritten many sections in this chapter to provide additional explanation and added context to the data-mining methods introduced.
- **New Material in Predictive Data Mining Chapters.** We have divided the previous single chapter on predictive data mining methods into two different chapters: Chapter 10 covers regression data mining models, and Chapter 11 covers classification data mining models. Splitting this chapter into two chapters allows us to include additional methods and more thoroughly explain the methods and output. Chapter 10 focuses on predicting quantitative outcomes with *k*-nearest neighbors regression, regression trees, and neural network regression. Chapter 11 focuses on predicting binary categorical outcomes with logistic regression, *k*-nearest neighbors classification, classification trees, and neural network classification.
- **New Online Appendixes for Using Python-Based Orange for Data Mining.** We now include online appendixes that introduce the software package Orange for descriptive and predictive data-mining models. Orange is an open-source machine learning and data visualization software package built using Python. It provides an easy-to-use, yet powerful, workflow-based approach to building analytics models. The use of Python by

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

analytics professionals has increased greatly, and Orange provides an excellent introduction to using Python in an easy-to-learn environment. These appendixes include practice problems for students to solve using Orange.

- **Updates to R Appendixes for Data Mining Chapters.** We continue to include online appendixes for using the popular open-source software package R for descriptive and predictive data mining. We have revised these online appendixes to focus on using script R commands in RStudio rather than the Rattle interface. We have found the use of script R through RStudio to be more robust and more stable than the use of Rattle. To facilitate the teaching of this material, we include complete R script files for all examples used throughout the R appendixes included for this this textbook as well as detailed step-by-step instructions.
- **Practice Problems in Online Appendixes for Orange and R.** Each online appendix for using Orange or R now includes practice problems that can be solved using these popular open-source software packages. The practice problems include specific hints for how to solve these problems using each software. Problem solutions, including the necessary code files, using these software packages are available to instructors for all practice problems.
- **Additional Coverage of Descriptive Analytics and Data Visualization.** We have added additional coverage in Chapters 2 and 3 to provide more depth of our coverage of descriptive analytics methods and data visualization. In Chapter 2, we have rewritten the explanation of histograms and we have added a discussion of frequency polygons to supplement the coverage of histograms as a way of exploring distributions of data. Chapter 3's coverage of data visualization now includes a more comprehensive discussion of best practices in data visualization including an explanation of the use of preattentive attributes and the data-ink ratio to create effective tables and charts. We have also rearranged the chapter and added coverage of table lens, waterfall charts, stock charts, choropleth maps, and cartograms for data visualization.
- **New Material for Optimization.** We have expanded our coverage of optimization models. Chapter 14 on Linear Optimization Models now includes additional models for transportation problems, diet problems, and assignment problems. Chapter 16 on nonlinear optimization models contains a new section on heuristic optimization using Excel's Evolutionary Solver method that can be used to solve complex optimization models. We have also added online appendixes for solving optimization models in R.
- **Increased Size of Data Sets.** We have increased the size of many data sets in Chapter 8 on linear regression and Chapter 9 on time series analysis and forecasting. These larger data sets provide better representations of realistic data sets that are encountered in practice, and they are more useful for conducting inference and examining the underlying assumptions made in regression and time series analysis.
- **Learning Objectives.** We have added Learning Objectives (LOs) to the beginning of each chapter. These LOs explain the key concepts that are covered in each chapter. The LOs are also mapped onto each end-of-chapter problem and case so instructors can easily identify which LOs are covered by each problem.
- **New End-of-Chapter Problems and Cases.** The fifth edition of this textbook includes over 175 new problems and 6 new cases. We have added problems to Chapter 1 so that now each chapter in the textbook includes end-of-chapter problems. We have divided the end-of-chapter problems in Chapters 6, 10, and 11 covering data mining topics into conceptual problems and software application problems. Conceptual problems do not require dedicated software to answer, and we have added many new conceptual problems to each data mining chapter. Software application problems require the use of dedicated software, and allow students to apply the tools covered in these chapters. The software application problems in these chapters are also provided in the online appendixes for the use of Orange and R for data mining applications. The online appendix version of these

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

problems include specific hints for how to solve the problems in the corresponding software. As we have done in past editions, Excel solution files are available to instructors for many of the problems and cases that require the use of Excel. For software application problems that require the use of software in the data-mining chapters (Chapters 6, 10, and 11), we include solutions for both Orange and R.

Continued Features and Pedagogy

In the fifth edition of this textbook, we continue to offer all of the features that have been successful in the previous editions. Some of the specific features that we use in this textbook are listed below.

- **Integration of Microsoft Excel:** Excel has been thoroughly integrated throughout this textbook. For many methodologies, we provide instructions for how to perform calculations both by hand and with Excel. In other cases where realistic models are practical only with the use of a spreadsheet, we focus on the use of Excel to describe the methods to be used. Excel instructions have been fully updated throughout the textbook to match the latest versions of Excel most likely to be used by students. The textbook assumes the use of desktop versions of Excel for most problems and examples, but the accompanying online Appendix D covers the basics of the use of Microsoft Excel Online.
- **Notes and Comments:** At the end of many sections, we provide Notes and Comments to give the student additional insights about the methods presented in that section. These insights include comments on the limitations of the presented methods, recommendations for applications, and other matters. Additionally, margin notes are used throughout the textbook to provide additional insights and tips related to the specific material being discussed.
- **Analytics in Action:** Each chapter contains an Analytics in Action article. These articles present interesting examples of the use of business analytics in practice. The examples are drawn from many different organizations in a variety of areas including healthcare, finance, manufacturing, marketing, and others.
- **DATAfiles and MODELfiles:** All data sets used as examples and in student exercises are also provided online on the companion site as files available for download by the student. DATAfiles are files that contain data needed for the examples and problems given in the textbook. Typically, the DATAfiles are in .xlsx format for Excel or .csv format for import into other software packages. MODELfiles contain additional modeling features such as extensive use of Excel formulas or the use of Excel Solver, script files for R, or workflow models for Orange.
- **Problems and Cases:** Each chapter, now including Chapter 1, contains an extensive selection of problems to help the student master the material presented in that chapter. The problems vary in difficulty and most relate to specific examples of the use of business analytics in practice. Answers to selected even-numbered problems are provided in an online supplement for student access. With the exception of Chapter 1, each chapter also includes at least one in-depth case study that connects many of the different methods introduced in the chapter. The case studies are designed to be more open-ended than the chapter problems, but enough detail is provided to give the student some direction in solving the cases. We continue to include two cases at the end of the textbook that require the use of material from multiple chapters in the text to better illustrate how concepts from different chapters relate to each other.

MindTap

MindTap is a customizable digital course solution that includes an interactive eBook, autograded exercises from the textbook, algorithmic practice problems with solutions feedback, Excel Online problems, Exploring Analytics visualizations, Adaptive Test Prep, videos, and

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

more! MindTap is also where instructors and users can find the online appendixes for R and Orange. All of these materials offer students better access to resources to understand the materials within the course. For more information on MindTap, please contact your Cengage representative.

WebAssign

Prepare for class with confidence using WebAssign from Cengage. This online learning platform fuels practice, so students can truly absorb what you learn—and are better prepared come test time. Videos, Problem Walk-Throughs, and End-of-Chapter problems and cases with instant feedback help them understand the important concepts, while instant grading allows you and them to see where they stand in class. WebAssign is also where instructors and users can find the online appendixes for R and Orange. Class Insights allows students to see what topics they have mastered and which they are struggling with, helping them identify where to spend extra time. Study Smarter with WebAssign.

Instructor and Student Resources

Additional instructor and student resources for this product are available online. Instructor assets include a Solutions and Answers Guide, Instructor's Manual, PowerPoint® slides, and a test bank powered by Cognero®. Prepared by the authors, solutions for software application problems and cases in chapters 6, 10, and 11 are available using both R and Orange. Student assets include DATAfiles and MODELfiles that accompany the chapter examples and problems as well as solutions to selected even-numbered chapter problems. Sign up or sign in at www.cengage.com to search for and access this product and its online resources.

Acknowledgments

We would like to acknowledge the work of reviewers and users who have provided comments and suggestions for improvement of this text. Thanks to:

Rafael Becerril Arreola
University of South Carolina

Matthew D. Bailey
Bucknell University

Phillip Beaver
University of Denver

M. Khurrum S. Bhutta
Ohio University

Paolo Catasti
Virginia Commonwealth University

Q B. Chung
Villanova University

Elizabeth A. Denny
University of Kentucky

Mike Taein Eom
University of Portland

Yvette Njan Essounga
Fayetteville State University

Lawrence V. Fulton
Texas State University

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Tom Groleau
Carthage College

James F. Hoelscher
Lincoln Memorial University

Eric Huggins
Fort Lewis College

Faizul Huq
Ohio University

Marco Lam
York College of Pennsylvania

Thomas Lee
University of California, Berkeley

Roger Myerson
Northwestern University

Ram Pakath
University of Kentucky

Susan Palocsay
James Madison University

Andy Shogan
University of California, Berkeley

Dothan Truong
Embry-Riddle Aeronautical University

Kai Wang
Wake Technical Community College

Ed Wasil
American University

Ed Winkofsky
University of Cincinnati

A special thanks goes to our associates from business and industry who supplied the Analytics in Action features. We recognize them individually by a credit line in each of the articles. We are also indebted to our senior product manager, Aaron Arnsparger; our Senior Content Manager, Breanna Holmes; senior learning designer, Brandon Foltz; subject matter expert, Deborah Cernauskas; digital project manager, Andrew Southwell; and our project manager at MPS Limited, Shreya Tiwari, for their editorial counsel and support during the preparation of this text.

Jeffrey D. Camm
James J. Cochran
Michael J. Fry
Jeffrey W. Ohlmann

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 1

Introduction to Business Analytics

Contents

1.1 Decision Making

1.2 Business Analytics Defined

1.3 A Categorization of Analytical Methods and Models
- Descriptive Analytics
- Predictive Analytics
- Prescriptive Analytics

1.4 Big Data, the Cloud, and Artificial Intelligence
- Volume
- Velocity
- Variety
- Veracity

1.5 Business Analytics in Practice
- Accounting Analytics
- Financial Analytics
- Human Resource (HR) Analytics
- Marketing Analytics
- Health Care Analytics
- Supply Chain Analytics
- Analytics for Government and Nonprofits
- Sports Analytics
- Web Analytics

1.6 Legal and Ethical Issues in the Use of Data and Analytics

Summary 15

Glossary 15

Problems 16

Available in the Cengage eBook:
- Appendix: Getting Started with R and Rstudio
- Appendix: Basic Data Manipulation with R

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Identify strategic, tactical, and operational decisions.

LO 2 Describe the steps in the decision-making process.

LO 3 Identify examples of descriptive, predictive, and prescriptive analytics.

LO 4 Describe applications of analytics for decision making.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In many situations, data are often abundant and can be used to guide decision-making. Suppose you apply for a loan for the first time. How does the bank assess the riskiness of the loan it might make to you? How does Amazon.com know which books and other products to recommend to you when you log in to their web site? How do airlines determine what price to quote to you when you are shopping for a plane ticket?

You may be applying for a loan for the first time, but millions of people around the world have applied for loans before. Many of these loan recipients have paid back their loans in full and on time, but some have not. The bank wants to know whether you are more like those who have paid back their loans or more like those who defaulted. By comparing your credit history, financial situation, and other factors to the vast database of previous loan recipients, the bank can effectively assess how likely you are to default on a loan.

Similarly, Amazon.com has access to data on millions of purchases made by customers on its web site. Amazon.com examines your previous purchases, the products you have viewed, and any product recommendations you have provided. Amazon.com then searches through its huge database for customers who are similar to you in terms of product purchases, recommendations, and interests. Once similar customers have been identified, their purchases form the basis of the recommendations given to you.

Prices for airline tickets are frequently updated. The price quoted to you for a flight between New York and San Francisco today could be very different from the price that will be quoted tomorrow. These changes happen because airlines use a pricing strategy known as revenue management. Revenue management works by examining vast amounts of data on past airline customer purchases and using these data to forecast future purchases. These forecasts are then fed into sophisticated optimization algorithms that determine the optimal price to charge for a particular flight and when to change that price. Revenue management has resulted in substantial increases in airline revenues.

This book is concerned with data-driven decision making and the use of analytical approaches in the decision-making process. Three developments spurred recent explosive growth in the use of analytical methods in business applications. First, technological advances have enabled the ability to track and store large amounts of data. Improved point-of-sale scanner technology, tracking of activity on the internet (e.g,. e-commerce and social networks), sensors on mechanical devices such as aircraft engines, automobiles, and farm machinery through the Internet-of-Things, and personal electronic devices produce incredible amounts of data for businesses. Naturally, businesses want to use these data to improve the efficiency and profitability of their operations, better understand their customers, price their products more effectively, and gain a competitive advantage. Second, ongoing research has resulted in numerous methodological developments to extract knowledge from the data. Examples of these are advances in computational approaches to effectively handle and explore massive amounts of data, faster algorithms for optimization and simulation, and more effective approaches for visualizing data. Third, these methodological developments are paired with an explosion in computing power. Faster computer chips, parallel computing, and cloud computing (the remote use of hardware and software over the Internet) have enabled businesses to solve big problems more quickly and more accurately than ever before.

In summary, the availability of massive amounts of data, improvements in analytic methodologies, and substantial increases in computing power have all come together to result in a dramatic upsurge in the use of analytical methods in business and a reliance on the discipline that is the focus of this text: business analytics. The purpose of this text is to provide students with a sound conceptual understanding of the role that business analytics plays in the decision-making process and to provide a better understanding of the variety of applications in which analytical methods have been used successfully.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

1.1 Decision Making

It is the responsibility of managers to plan, coordinate, organize, and lead their organizations to better performance. Ultimately, managers' responsibilities require that they make strategic, tactical, or operational decisions. **Strategic decisions** involve higher-level issues concerned with the overall direction of the organization; these decisions define the organization's overall goals and aspirations for the future. Strategic decisions are usually the domain of higher-level executives and have a time horizon of three to five years. **Tactical decisions** concern how the organization should achieve the goals and objectives set by its strategy, and they are usually the responsibility of midlevel management. Tactical decisions usually span a year and thus are revisited annually or even every six months. **Operational decisions** affect how the firm is run from day to day; they are the domain of operations managers, who are the closest to the customer.

Consider the case of the Thoroughbred Running Company (TRC). Historically, TRC had been a catalog-based retail seller of running shoes and apparel. TRC sales revenues grew quickly as it changed its emphasis from catalog-based sales to Internet-based sales. Recently, TRC decided that it should also establish retail stores in the malls and downtown areas of major cities. This strategic decision will take the firm in a new direction that it hopes will complement its Internet-based strategy. TRC middle managers will therefore have to make a variety of tactical decisions in support of this strategic decision, including how many new stores to open this year, where to open these new stores, how many distribution centers will be needed to supply the new stores, and where to locate these distribution centers. Operations managers in the stores will need to make day-to-day decisions regarding, for instance, how many pairs of each model and size of shoes to order from the distribution centers and how to schedule their sales personnel's work time.

Regardless of the level within the firm, *decision making* can be defined as the following process:

1. Identify and define the problem.
2. Determine the criteria that will be used to evaluate alternative solutions.
3. Determine the set of alternative solutions.
4. Evaluate the alternatives.
5. Choose an alternative.

Step 1 of decision making, identifying and defining the problem, is the most critical. Only if the problem is well-defined, with clear metrics of success or failure (step 2), can a proper approach for solving the problem (steps 3 and 4) be devised. Decision making concludes with the choice of one of the alternatives (step 5).

There are a number of approaches to making decisions: tradition ("We've always done it this way"), intuition ("gut feeling"), and rules of thumb ("As the restaurant owner, I schedule twice the number of waiters and cooks on holidays"). The power of each of these approaches should not be underestimated. Managerial experience and intuition are valuable inputs to making decisions, but what if relevant data are available to help us make more-informed decisions? The vast amounts of data now generated and stored electronically provide tremendous opportunity for businesses to improve their profitability and service to their customers. How can managers convert these data into knowledge they can use to be more efficient and effective in managing their businesses?

1.2 Business Analytics Defined

What makes decision making difficult and challenging? Uncertainty is probably the number one challenge. If we knew how much the demand would be for our product, we could do a much better job of planning and scheduling production. If we knew exactly how long each step in a project would take to be completed, we could better predict the project's cost and completion date. If we knew how stocks would perform, investing would be a lot easier.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Another factor that makes decision making difficult is that we often face such an enormous number of alternatives that we cannot evaluate them all. What is the best combination of stocks to help me meet my financial objectives? What is the best product line for a company that wants to maximize its market share? How should an airline price its tickets so as to maximize revenue?

*Some firms and industries use the simpler term, **analytics**. Analytics is often thought of as a broader category than business analytics, encompassing the use of analytical techniques in the sciences and engineering as well. In this text, we use **business analytics** and **analytics** synonymously.*

Business analytics is the scientific process of transforming data into insight for making better decisions.[1] Business analytics is used for data-driven or fact-based decision making, which is often seen as more objective than other alternatives for decision making.

As we shall see, the tools of business analytics can aid decision making by creating insights from data, by improving our ability to more accurately forecast for planning, by helping us quantify risk, and by yielding better alternatives through analysis and optimization. A study based on a large sample of firms that was conducted by researchers at MIT's Sloan School of Management and the University of Pennsylvania concluded that firms guided by data-driven decision making have higher productivity and market value and increased output and profitability.[2]

1.3 A Categorization of Analytical Methods and Models

Business analytics can involve anything from simple reports to the most advanced optimization techniques (methods for finding the best course of action). Analytics is generally thought to comprise three broad categories of techniques: descriptive analytics, predictive analytics, and prescriptive analytics.

Descriptive Analytics

Descriptive analytics encompasses the set of techniques that describes what has happened in the past. Examples are data queries, reports, descriptive statistics, data visualization including data dashboards, unsupervised learning techniques from data mining, and basic spreadsheet models.

Appendix B, at the end of this book, describes how to use Microsoft Access to conduct data queries.

A **data query** is a request for information with certain characteristics from a database. For example, a query to a manufacturing plant's database might be for all records of shipments to a particular distribution center during the month of March. This query provides descriptive information about these shipments: the number of shipments, how much was included in each shipment, the date each shipment was sent, and so on. A report summarizing relevant historical information for management might be conveyed by the use of descriptive statistics (means, measures of variation, etc.) and data-visualization tools (tables, charts, and maps). Simple descriptive statistics and data-visualization techniques can be used to find patterns or relationships in a large database.

Data dashboards are collections of tables, charts, maps, and summary statistics that are updated as new data become available. Dashboards are used to help management monitor specific aspects of the company's performance related to their decision-making responsibilities. For corporate-level managers, daily data dashboards might summarize sales by region, current inventory levels, and other company-wide metrics; front-line managers may view dashboards that contain metrics related to staffing levels, local inventory levels, and short-term sales forecasts.

Data mining is the use of analytical techniques for better understanding patterns and relationships that exist in large data sets. Data mining includes **unsupervised learning** techniques which are descriptive methods that seek to identify patterns based on notions of similarity (cluster analysis) or correlation (association rules) in different types of data. For

[1]We adopt the definition of analytics developed by the Institute for Operations Research and the Management Sciences (INFORMS).

[2]E. Brynjolfsson, L. M. Hitt, and H. H. Kim, "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" Thirty-Second International Conference on Information Systems, Shanghai, China, December 2011.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

example, by processing text on social network platforms such as Twitter, similar customer comments can be grouped together into clusters to help Apple better understand how its customers are feeling about the Apple Watch.

Predictive Analytics

Predictive analytics consists of techniques that use models constructed from past data to predict the future or ascertain the impact of one variable on another. For example, past data on product sales may be used to construct a mathematical model to predict future sales. This model can factor in the product's growth trajectory and seasonality based on past patterns. A packaged-food manufacturer may use point-of-sale scanner data from retail outlets to help in estimating the lift in unit sales due to coupons or sales events. Survey data and past purchase behavior may be used to help predict the market share of a new product. All of these are applications of predictive analytics. Traditional statistical methods such as linear regression and time series analysis fall under the banner of predictive analytics.

Data mining includes **supervised learning** techniques which use past data to learn the relationship between an outcome variable of interest and a set of input variables. For example, a large grocery store chain might be interested in developing a targeted marketing campaign that offers a discount coupon on potato chips. By studying historical point-of-sale data, the store may be able to use data mining to predict which customers are the most likely to respond to an offer on discounted potato chips, by purchasing higher-margin items such as beer or soft drinks in addition to the chips, thus increasing the store's overall revenue.

Simulation involves the use of probability and statistics to construct a computer model to study the impact of uncertainty on a decision. For example, banks often use simulation to model investment and default risk in order to stress-test financial models. Simulation is also often used in the pharmaceutical industry to assess the risk of introducing a new drug.

Prescriptive Analytics

Prescriptive analytics differs from descriptive and predictive analytics in that **prescriptive analytics** indicates a course of action to take; that is, the output of a prescriptive model is a decision. Predictive models provide a forecast or prediction, but do not provide a decision. However, a forecast or prediction, when combined with a rule, becomes a prescriptive model. For example, we may develop a model to predict the probability that a person will default on a loan. If we create a rule that says if the estimated probability of default is more than 0.6, we should not award a loan, now the predictive model, coupled with the rule is prescriptive analytics. These types of prescriptive models that rely on a rule or set of rules are often referred to as **rule-based models**.

Other examples of prescriptive analytics are portfolio models in finance, supply network design models in operations, and price-markdown models in retailing. Portfolio models use historical investment return data to determine which mix of investments will yield the highest expected return while controlling or limiting exposure to risk. Supply-network design models provide plant and distribution center locations that will minimize costs while still meeting customer service requirements. Given historical data, retail price markdown models yield revenue-maximizing discount levels and the timing of discount offers when goods have not sold as planned. All of these models are known as **optimization models**, that is, models that give the best decision subject to the constraints of the situation.

Another type of modeling in the prescriptive analytics category is **simulation optimization** which combines the use of probability and statistics to model uncertainty with optimization techniques to find good decisions in highly complex and highly uncertain settings. Finally, the techniques of **decision analysis** can be used to develop an optimal strategy when a decision maker is faced with several decision alternatives and an uncertain set of future events. Decision analysis also employs **utility theory**, which assigns values to outcomes based on the decision maker's attitude toward risk, loss, and other factors.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 1.1 Coverage of Business Analytics Topics in This Text

Chapter	Title	Descriptive	Predictive	Prescriptive
1	Introduction	●	●	●
2	Descriptive Statistics	●		
3	Data Visualization	●		
4	Data Wrangling	●		
5	Probability: An Introduction to Modeling Uncertainty	●		
6	Descriptive Data Mining	●		
7	Statistical Inference	●		
8	Linear Regression		●	
9	Time Series & Forecasting		●	
10	Predictive Data Mining: Regression Tasks		●	
11	Predictive Data Mining: Classification Tasks		●	
12	Spreadsheet Models	●	●	●
13	Monte Carlo Simulation		●	●
14	Linear Optimization Models			●
15	Integer Optimization Models			●
16	Nonlinear Optimization Models			●
17	Decision Analysis			●

In this text, we cover all three areas of business analytics: descriptive, predictive, and prescriptive. Table 1.1 shows how the chapters cover the three categories.

1.4 Big Data, the Cloud, and Artificial Intelligence

On any given day, Google receives over 8.5 billion searches, WhatsApp users exchange up to 65 billion messages, and Internet users generate about 2.5 quintillion bytes of data.[3] It is through technology that we have truly been thrust into the data age. Because data can now be collected electronically, the available amounts of it are staggering. The Internet, cell phones, retail checkout scanners, surveillance video, and sensors on everything from aircraft to cars to bridges allow us to collect and store vast amounts of data in real time.

In the midst of all of this data collection, the term *big data* has been created. There is no universally accepted definition of big data. However, probably the most accepted and most general definition is that **big data** is any set of data that is too large or too complex to be handled by standard data-processing techniques and typical desktop software. IBM describes the phenomenon of big data through the four Vs: volume, velocity, variety, and veracity, as shown in Figure 1.1.

Volume

Because data are collected electronically, we are able to collect more of it. To be useful, these data must be stored, and this storage has led to vast quantities of data. Many companies now store in excess of 100 terabytes of data (a terabyte is 1,024 gigabytes).

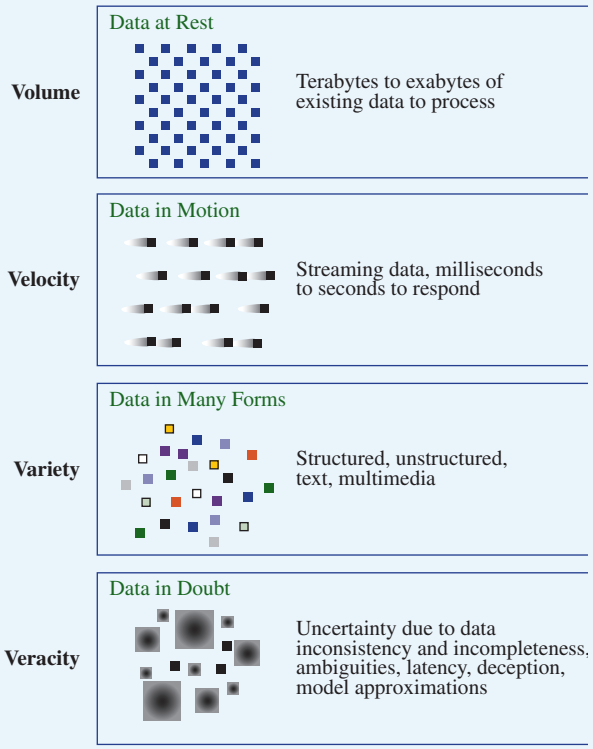
Velocity

Real-time capture and analysis of data present unique challenges both in how data are stored and the speed with which those data can be analyzed for decision making. For

[3]https://techjury.net/blog/big-data-statistics/#gref

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 1.1 The Four Vs of Big Data



Source: IBM.

example, the New York Stock Exchange collects 1 terabyte of data in a single trading session, and having current data and real-time rules for trades and predictive modeling are important for managing stock portfolios.

Variety

In addition to the sheer volume and speed with which companies now collect data, more complicated types of data are now available and are proving to be of great value to businesses. Text data are collected by monitoring what is being said about a company's products or services on social media platforms such as Twitter. Audio data are collected from service calls (on a service call, you will often hear "this call may be monitored for quality control"). Video data collected by in-store video cameras are used to analyze shopping behavior. Analyzing information generated by these nontraditional sources is more complicated in part because of the processing required to transform the data into a numerical form that can be analyzed.

Veracity

Veracity has to do with how much uncertainty is in the data. For example, the data could have many missing values, which makes reliable analysis a challenge. Inconsistencies in units of measure and the lack of reliability of responses also increase the complexity of the data.

Businesses have realized that understanding big data can lead to a competitive advantage. Although big data represents opportunities, it also presents challenges in terms of data storage and processing, security, and available analytical talent.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The four Vs indicate that big data creates challenges in terms of how these complex data can be captured, stored, processed, secured, and then analyzed. Traditional databases typically assume that data fit into nice rows and columns, but that is not always the case with big data. Also, the sheer volume (the first V) often means that it is not possible to store all of the data on a single computer. This has led to technologies such as **Hadoop**—an open-source programming environment that supports big data processing through distributed storage and distributed processing on clusters of computers. Essentially, Hadoop provides a divide-and-conquer approach to handling massive amounts of data, dividing the storage and processing over multiple computers. **MapReduce** is a programming model used within Hadoop that performs the two major steps for which it is named: the map step and the reduce step. The map step divides the data into manageable subsets and distributes it to the computers in the cluster (often termed nodes) for storing and processing. The reduce step collects answers from the nodes and combines them into an answer to the original problem. Technologies such as Hadoop and MapReduce, paired with relatively inexpensive computer power, enable cost-effective processing of big data; otherwise, in some cases, processing might not even be possible.

The massive amounts of available data have led to numerous innovations that help make the data useful for decision making. **Cloud computing**, also known simply as "the cloud," refers to the use of data and software on servers housed external to an organization via the internet. The cloud has made storing and processing massive amounts of data feasible and cost effective. Companies pay a fee to store their big data and use software stored on servers with cloud alternatives such as Amazon Web Services (AWS), Microsoft Azure, and Oracle Cloud.

Storing so much data has raised concerns about the security. Medical records, bank account information, and credit card transactions, for example, are all highly confidential and must be protected from computer hackers. **Data security**, the protection of stored data from destructive forces or unauthorized users, is of critical importance to companies. For example, credit card transactions are potentially very useful for understanding consumer behavior, but compromise of these data could lead to unauthorized use of the credit card or identity theft. Companies such as Target, Anthem, JPMorgan Chase, Yahoo!, Facebook, Marriott, Equifax, and Home Depot have faced major data breaches costing millions of dollars.

In addition to increased interest in cloud computing and data security, big data has accelerated the development of applications of artificial intelligence. Broadly speaking, **artificial intelligence** (AI) is the use of big data and computers to make decisions that in the past would have required human intelligence. Often, AI software mimics the way we understand the human brain functions. AI uses many of the techniques of analytics, but often in real time, with massive amounts of data required to "train" algorithms, to complete an automated task. Applications of AI include for example, facial recognition for security checkpoints and self-driving vehicles.

The complexities of the big data have increased the demand for analysts, but a shortage of qualified analysts has made hiring more challenging. More companies are searching for **data scientists**, who know how to effectively process and analyze massive amounts of data because they are well trained in both computer science and statistics. Next we discuss three examples of how companies are collecting big data for competitive advantage.

Kroger Understands Its Customers Kroger is the largest retail grocery chain in the United States. It sends over 11 million pieces of direct mail to its customers each quarter. The quarterly mailers each contain 12 coupons that are tailored to each household based on several years of shopping data obtained through its customer loyalty card program. By collecting and analyzing consumer behavior at the individual household level, and better matching its coupon offers to shopper interests, Kroger has been able to realize a far higher redemption rate on its coupons. In the six-week period following distribution of the mailers, over 70% of households redeem at least one coupon, leading to an estimated coupon revenue of $10 billion for Kroger. (Source: *Forbes.com*)

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

MagicBand at Disney The Walt Disney Company offers a wristband to visitors to its Orlando, Florida, Disney World theme park. Known as the MagicBand, the wristband contains technology that can transmit more than 40 feet and can be used to track each visitor's location in the park in real time. The band can link to information that allows Disney to better serve its visitors. For example, prior to the trip to Disney World, a visitor might be asked to fill out a survey on their birth date and favorite rides, characters, and restaurant table type and location. This information, linked to the MagicBand, can allow Disney employees using smartphones to greet you by name as you arrive, offer you products they know you prefer, wish you a happy birthday, have your favorite characters show up as you wait in line or have lunch at your favorite table. The MagicBand can be linked to your credit card, so there is no need to carry cash or a credit card. And during your visit, your movement throughout the park can be tracked and the data can be analyzed to better serve you during your next visit to the park. (Source: *Wired.com*)

Coca-Cola Freestyle Gives Consumers Their Own Personal Soft Drinks Coca-Cola, the largest beverage company in the world, sells its products in over 200 countries. One of Coca-Cola's innovations is Coca-Cola Freestyle, a touch screen self-service soda dispenser that can be found in numerous restaurants such as Subway and Burger King. Coca-Cola Freestyle allows the customer to create their own customized drink by combining existing flavors. For example, a customer might combine Sprite with Hi-C Orange to create a flavor that is not available in stores. But Coca-Cola Freestyle is not only an innovation that better serves customers through mass customization—the machines are also a goldmine of customer preference data. The Freestyle machines collect data on mixes that consumers around the world choose. These data are analyzed to better understand customer preferences and how they differ by country and region. Whereas in the past, marketing analysts would need to rely on expensive focus groups and multiple rounds of market testing to help develop new products, Freestyle machines directly provide data on consumer preferences. For example, a new bottled product, Cherry Sprite, was launched when Freestyle data indicated that demand for that combination of flavors would be strong in the United States. (Source: *buzzfeednews.com*)

Although big data is clearly one of the drivers for the strong demand for analytics, it is important to understand that, in some sense, big data issues are a subset of analytics. Many very valuable applications of analytics do not involve big data, but rather traditional data sets that are very manageable by traditional database and analytics software. The key to analytics is that it provides useful insights and better decision making using the data that are available—whether those data are "big" or not.

1.5 Business Analytics in Practice

Business analytics involves tools as simple as reports and graphs to those that are as sophisticated as optimization, data mining, and simulation. In practice, companies that apply analytics often follow a trajectory similar to that shown in Figure 1.2. Organizations start with basic analytics in the lower left. As they realize the advantages of these analytic techniques, they often progress to more sophisticated techniques in an effort to reap the derived competitive advantage. Therefore, predictive and prescriptive analytics are sometimes referred to as **advanced analytics**. Not all companies reach that level of usage, but those that embrace analytics as a competitive strategy often do.

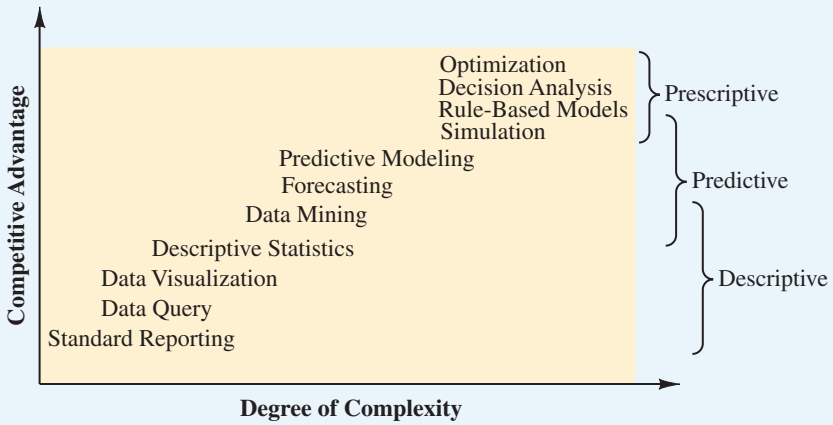
Analytics has been applied in virtually all sectors of business and government. Organizations such as Procter & Gamble, IBM, UPS, Netflix, Amazon.com, Google, the Internal Revenue Service, and General Electric have embraced analytics to solve important problems or to achieve a competitive advantage. In this section, we briefly discuss some of the types of applications of analytics by application area.

Accounting Analytics

Applications of analytics in accounting are numerous and growing. For example, budget planning relies heavily on analytics. To construct a budget, predictive analytics in the form

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 1.2 The Spectrum of Business Analytics



Source: Adapted from SAS.

forecasts are usually required for estimates of market demand for goods and services, exchange and inflation rates, and cost projections to name a few. Furthermore, risk analysis, usually through simulation, helps ensure that the budget is robust in the face of uncertainty. Also, the use of data dashboards to monitor the financial health of companies over time is quite common.

By analyzing financial records and financial transaction data, forensic accounting investigates incidents such as fraud, bribery, embezzlement, and money laundering. Descriptive and predictive analytics can be used to detect unusual patterns in transactions and to find outlier data that could be an indication of illegal behavior.

Financial Analytics

The financial services sector relies heavily on descriptive, predictive, and prescriptive analytics. Descriptive analytics, through the use of data visualization, is used to monitor financial performance, including stock returns, trading volumes, and measures of market return and volatility. Predictive models are used to forecast financial performance, to assess the risk of investment portfolios and projects, and to construct financial instruments such as derivatives. Prescriptive models are used to construct optimal portfolios of investments, to allocate assets, and to create optimal capital budgeting plans.

Human Resource (HR) Analytics

A relatively new area of application for analytics is the management of an organization's human resources. The HR function is charged with ensuring that the organization (1) has the mix of skill sets necessary to meet its needs, (2) is hiring the highest-quality talent and providing an environment that retains it, and (3) achieves its organizational diversity goals. Google refers to its HR Analytics function as "people analytics." Google has analyzed substantial data on their own employees to determine the characteristics of great leaders, to assess factors that contribute to productivity, and to evaluate potential new hires. Google also uses predictive analytics to continually update their forecast of future employee turnover and retention.

Marketing Analytics

Marketing is one of the fastest-growing areas for the application of analytics. A better understanding of consumer behavior through the use of scanner data and data generated from social media has led to an increased interest in marketing analytics. As a result,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

descriptive, predictive, and prescriptive analytics are all heavily used in marketing. A better understanding of consumer behavior through analytics leads to the better use of advertising budgets, more effective pricing strategies, improved forecasting of demand, improved product-line management, and increased customer satisfaction and loyalty. Predictive models and optimization are used to better align advertising to specific target audiences, making those marketing efforts much more effective and efficient. Likewise, search engine optimization increases the effectiveness of product searches, which leads to increased sales from e-commerce customers.

Sentiment analysis allows companies to monitor the reactions of their customers to the goods and services they offer and adjust their service and product offerings based on that data. Through sentiment analysis, "the voice of the customer" is more powerful than ever before.

Finally, the use of analytics to determine prices to charge for products and to optimally mark down the price of seasonal goods over time have helped increase profitability in the retail space. Revenue management adjusts prices for perishable goods and services based on changes in demand over time. This has led to increased profitability for airlines, hotels, rental car companies, and sports franchises.

Health Care Analytics

The use of analytics in health care is on the increase because of pressure to simultaneously control costs and provide more effective treatment. Descriptive, predictive, and prescriptive analytics are used to improve patient, staff, and facility scheduling; patient flow; purchasing; and inventory control. Predictive analytics is used to predict which patients are most likely to not adhere to their required medications so that proactive reminders can be sent to help these patients. Prescriptive analytics has been used directly to provide better patient outcomes. For example, researchers at Georgia Tech helped develop approaches that optimally place radiation seeds to treat prostate cancer so that the cancer is eradicated while doing minimal damage to healthy tissue.

Supply Chain Analytics

The core service of companies such as UPS and FedEx is the efficient delivery of goods, and analytics has long been used to achieve efficiency. The optimal sorting of goods, vehicle and staff scheduling, and vehicle routing are all key to profitability for logistics companies such as UPS and FedEx.

While supply chain analytics has historically focused almost exclusively on efficiency, the supply chain problems caused by the COVID-19 pandemic and world conflicts have expanded the use of analytics to increase the resiliency of the supply chain. Monitoring and quantifying the riskiness of the supply chain and taking data-driven action to ensure the supply chain can support its mission over a robust set of scenarios has taken on new importance. Descriptive analytics is being used to more closely monitor supply chain performance, predictive analytics quantifies risk, and prescriptive analytics is used with scenario analysis to prepare supply chain solutions that can handle a high degree of disruption.

Analytics for Government and Nonprofits

Government agencies and other nonprofits have used analytics to drive out inefficiencies and increase the effectiveness and accountability of programs. Indeed, much of advanced analytics has its roots in the U.S. and English military dating back to World War II. Today, the use of analytics in government is becoming pervasive in everything from elections to tax collection. The U.S. Internal Revenue Service has used data mining to identify patterns that distinguish questionable annual personal income tax filings. In one application, the IRS compares its data on individual taxpayers with data received from banks on mortgage payments made by those taxpayers. When taxpayers report a mortgage payment that is unrealistically high relative to their reported taxable income, they are flagged as possible underreporters of taxable income. The filing is then further scrutinized and may trigger an audit.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Likewise, nonprofit agencies have used analytics to ensure their effectiveness and accountability to their donors and clients. Descriptive and predictive analytics are used to monitor agency performance as well as track donor behavior and forecast donations. Fundraising analytics, using techniques from data mining, are used to identify potential donors and minimize donor attrition. Optimization is used to allocate scarce resources in the capital budgeting process.

Sports Analytics

The use of analytics in sports has gained considerable notoriety since 2003 when renowned author Michael Lewis published *Moneyball*. Lewis' book tells the story of how the Oakland Athletics used an analytical approach to player evaluation in order to assemble a competitive team with a limited budget. The use of analytics for player evaluation and on-field strategy is now common, especially in professional sports. Professional sports teams use analytics to assess players for the amateur drafts and to decide how much to offer players in contract negotiations; and teams use analytics to assist with on-field decisions such as which pitchers to use in various games of a Major League Baseball playoff series.

The use of analytics for off-the-field business decisions is also increasing rapidly. Ensuring customer satisfaction is important for any company, and fans are the customers of sports teams. Based on fan survey data, the Cleveland Guardians professional baseball team used a type of predictive modeling known as conjoint analysis to design its premium seating offerings at Progressive Field. Using prescriptive analytics, franchises across several major sports dynamically adjust ticket prices throughout the season to reflect the relative attractiveness and potential demand for each game.

Web Analytics

Web analytics is the analysis of online activity, which includes, but is not limited to, visits to web sites and social media sites such as Facebook and LinkedIn. Web analytics has huge implications for promoting and selling products and services via the Internet. Leading companies apply descriptive and advanced analytics to data collected in online experiments to determine the best way to configure web sites, position ads, and utilize social networks for the promotion of products and services. Online experimentation involves exposing various subgroups to different versions of a web site and tracking the results. Because of the massive pool of Internet users, experiments can be conducted without risking the disruption of the overall business of the company. Such experiments are proving to be invaluable because they enable the company to use trial-and-error in determining statistically what makes a difference in their web site traffic and sales.

1.6 Legal and Ethical Issues in the Use of Data and Analytics

With the advent of big data and the dramatic increase in the use of analytics and data science to improve decision making, increased attention has been paid to ethical concerns around data privacy and the ethical use of models based on data.

As businesses routinely collect data about their customers, they have an obligation to protect the data and to not misuse that data. Clients and customers have an obligation to understand the trade-offs between allowing their data to be collected and used, and the benefits they accrue from allowing a company to collect and use that data. For example, many companies have loyalty cards that collect data on customer purchases. In return for the benefits of using a loyalty card, typically discounted prices, customers must agree to allow the company to collect and use the data on purchases. An agreement must be signed between the customer and the company, and the agreement must specify what data will be collected and how it will be used. For example, the agreement might say that all scanned purchases will be collected with the date, time, location, and card number, but that the company agrees to only use that data internally to the company and to not give or sell that data to outside firms or individuals.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The company then has an ethical obligation to uphold that agreement and make every effort to ensure that the data are protected from any type of unauthorized access. Unauthorized access of data is known as a data breach. Data breaches are a major concern for all companies in the digital age. A study by IBM and the Ponemon Institute estimated that the average cost of a data breach is $3.86 million.

Data privacy laws are designed to protect individuals' data from being used against their wishes. One of the strictest data privacy laws is the General Data Protection Regulation (GDPR) which went into effect in the European Union in May 2018. The law stipulates that the request for consent to use an individual's data must be easily understood and accessible, the intended uses of the data must be specified, and it must be easy to withdraw consent. The law also stipulates that an individual has a right to a copy of their data and the right "to be forgotten," that is, the right to demand that their data be erased. It is the responsibility of analytics professionals, indeed, anyone who handles or stores data, to understand the laws associated with the collection, storage, and use of individuals' data.

Ethical issues that arise in the use of data and analytics are just as important as the legal issues. Analytics professionals have a responsibility to behave ethically, which includes protecting data, being transparent about the data and how it was collected, and what it does and does not contain. Analysts must be transparent about the methods used to analyze the data and any assumptions that have to be made for the methods used. Finally, analysts must provide valid conclusions and understandable recommendations to their clients.

Intentionally using biased data to achieve a goal is inherently unethical. Misleading a client by misrepresenting results is clearly unethical. For example, consider the case of an airline that runs an advertisement that "it is preferred by 84% of business fliers to Chicago." Such a statement is valid if the airline randomly surveyed business fliers across all airlines with a destination of Chicago. But, if for convenience, the airline surveyed only its own customers, the survey would be biased, and the claim would be misleading because fliers on other airlines were not surveyed. Indeed, if anything, the only conclusion one can legitimately draw from the biased sample of its own customers would be that 84% of that airlines' own customers preferred that airline and 16% of its own customers actually preferred another airline!

In the book, *Weapons of Math Destruction*, author Cathy O'Neil discusses how algorithms and models can be unintentionally biased. For example, consider an analyst who is building a credit risk model for awarding loans. The location of the home of the applicant might be a variable that is correlated with other variables such as income and ethnicity. Income is perhaps a relevant variable for determining the amount of a loan, but ethnicity is not. A model using home location could therefore lead to unintentional bias in the credit risk model. It is the analysts' responsibility to make sure this type of model bias and data bias do not become a part of the model.

Researcher and opinion writer Zeynep Tufecki examines so-called "unintended consequences" of analytics and artificial intelligence, and particularly of machine learning and recommendation engines. Many Internet sites that use recommendation engines often suggest more extreme content, in terms of political views and conspiracy theories, to users based on their past viewing history. Tufecki and others theorize that this is because the machine learning algorithms being used have identified that more extreme content increases users' viewing time on the site, which is often the objective function being maximized by the machine learning algorithm. Therefore, while it is not the intention of the algorithm to promote more extreme views and disseminate false information, this may be the unintended consequence of using a machine learning algorithm that maximizes users' viewing time on the site. Analysts and decision makers must be aware of potential unintended consequences of their models, and they must decide how to react to these consequences once they are discovered.

Several organizations, including the American Statistical Association (ASA) and the Institute for Operations Research and the Management Sciences (INFORMS), provide ethical guidelines for analysts. In their "Ethical Guidelines for Statistical Practice," the ASA uses the term *statistician* throughout, but states that this "includes all practitioners

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

of statistics and quantitative sciences—regardless of job title or field of degree—comprising statisticians at all levels of the profession and members of other professions who utilize and report statistical analyses and their applications." Their guidelines state that "Good statistical practice is fundamentally based on transparent assumptions, reproducible results, and valid interpretations." The report contains 52 guidelines organized into 8 topic areas. We encourage you to read and familiarize yourself with these guidelines.

INFORMS is a professional society focused on operations research and the management sciences, including analytics. INFORMS offers an analytics certification called CAP—certified analytics professional. All candidates for CAP are required to comply with the code of ethics/conduct provided by INFORMS. The INFORMS CAP guidelines state, "In general, analytics professionals are obliged to conduct their professional activities responsibly, with particular attention to the values of consistency, respect for individuals, autonomy of all, integrity, justice, utility and competence." INFORMS also offers a set of Ethics Guidelines for its members, which covers ethical behavior for analytics professionals in three domains: Society, Organizations (businesses, government, nonprofit organization, and universities), and the Profession (operations research and analytics). As these guidelines are fairly easy to understand and at the same time fairly comprehensive, we list them here in Table 1.2 and encourage you as a user/provider of analytics to make them your guiding principles.

Table 1.2 INFORMS Ethics Guidelines

Relative to Society

Analytics professionals should aspire to be:

- **Accountable** for their professional actions and the impact of their work.
- **Forthcoming** about their assumptions, interests, sponsors, motivations, limitations, and potential conflicts of interest.
- **Honest** in reporting their results, even when they fail to yield the desired outcome.
- **Objective** in their assessments of facts, irrespective of their opinions or beliefs.
- **Respectful** of the viewpoints and the values of others.
- **Responsible** for undertaking research and projects that provide positive benefits by advancing our scientific understanding, contributing to organizational improvements, and supporting social good.

Relative to Organizations

Analytics professionals should aspire to be:

- **Accurate** in our assertions, reports, and presentations.
- **Alert** to possible unintended or negative consequences that our results and recommendations may have on others.
- **Informed** of advances and developments in the fields relevant to our work.
- **Questioning** of whether there are more effective and efficient ways to reach a goal.
- **Realistic** in our claims of achievable results, and in acknowledging when the best course of action may be to terminate a project.
- **Rigorous** by adhering to proper professional practices in the development and reporting of our work.

Relative to the Profession

Analytics professionals should aspire to be:

- **Cooperative** by sharing best practices, information, and ideas with colleagues, young professionals, and students.
- **Impartial** in our praise or criticism of others and their accomplishments, setting aside personal interests.
- **Inclusive** of all colleagues, and rejecting discrimination and harassment in any form.
- **Tolerant** of well-conducted research and well-reasoned results, which may differ from our own findings or opinions.
- **Truthful** in providing attribution when our work draws from the ideas of others.
- **Vigilant** by speaking out against actions that are damaging to the profession

Source: INFORMS Ethics Guidelines.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary

This introductory chapter began with a discussion of decision making. Decision making can be defined as the following process: (1) identify and define the problem, (2) determine the criteria that will be used to evaluate alternative solutions, (3) determine the set of alternative solutions, (4) evaluate the alternatives, and (5) choose an alternative. Decisions may be strategic (high level, concerned with the overall direction of the business), tactical (mid-level, concerned with how to achieve the strategic goals of the business), or operational (day-to-day decisions that must be made to run the company).

Uncertainty and an overwhelming number of alternatives are two key factors that make decision making difficult. Business analytics approaches can assist by identifying and mitigating uncertainty and by prescribing the best course of action from a very large number of alternatives. In short, business analytics can help us make better-informed decisions.

There are three categories of analytics: descriptive, predictive, and prescriptive. Descriptive analytics describes what has happened and includes tools such as reports, data visualization, data dashboards, descriptive statistics, and unsupervised learning techniques from data mining. Predictive analytics consists of techniques that use past data to predict future events or ascertain the impact of one variable on another. These techniques include regression, supervised learning techniques from data mining, forecasting, and simulation. Prescriptive analytics uses data to determine a course of action. This class of analytical techniques includes rule-based models, simulation, decision analysis, and optimization. Descriptive and predictive analytics can help us better understand the uncertainty and risk associated with our decision alternatives. Predictive and prescriptive analytics, also often referred to as advanced analytics, can help us make the best decision when facing a myriad of alternatives.

Big data is a set of data that is too large or too complex to be handled by standard data-processing techniques or typical desktop software. Cloud computing has made storing and processing vast amounts of data more efficient and more cost effective. The increasing prevalence of big data is leading to an increase in the use of analytics and artificial intelligence. The Internet, retail scanners, and cell phones are making huge amounts of data available to companies, and these companies want to better understand these data. Business analytics helps them understand these data and use them to make better decisions.

We also discussed various application areas of analytics. Our discussion focused on accounting analytics, financial analytics, human resource analytics, marketing analytics, health care analytics, supply chain analytics, analytics for government and nonprofit organizations, sports analytics, and web analytics. However, the use of analytics is rapidly spreading to other sectors, industries, and functional areas of organizations. We concluded this chapter with a discussion of legal and ethical issues in the use of data and analytics, a topic that should be of great importance to all practitioners and consumers of analytics. Each remaining chapter in this text will provide a real-world vignette in which business analytics is applied to a problem faced by a real organization.

Glossary

Artificial intelligence (AI) The use of data and computers to make decisions that would have in the past required human intelligence.
Advanced analytics Predictive and prescriptive analytics.
Big data Any set of data that is too large or too complex to be handled by standard data-processing techniques and typical desktop software.
Business analytics The scientific process of transforming data into insight for making better decisions.
Cloud computing The use of data and software on severs housed external to an organization via the internet. Also known simply as "the cloud."
Data dashboard A collection of tables, charts, and maps to help management monitor selected aspects of the company's performance.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Data mining The use of analytical techniques for better understanding patterns and relationships that exist in large data sets.
Data query A request for information with certain characteristics from a database.
Data scientists Analysts trained in both computer science and statistics who know how to effectively process and analyze massive amounts of data.
Data security Protecting stored data from destructive forces or unauthorized users.
Decision analysis A technique used to develop an optimal strategy when a decision maker is faced with several decision alternatives and an uncertain set of future events.
Descriptive analytics Analytical tools that describe what has happened.
Hadoop An open-source programming environment that supports big data processing through distributed storage and distributed processing on clusters of computers.
MapReduce Programming model used within Hadoop that performs the two major steps for which it is named: the map step and the reduce step.
Operational decision A decision concerned with how the organization is run from day to day.
Optimization model A mathematical model that gives the best decision, subject to the situation's constraints.
Predictive analytics Techniques that use models constructed from past data to predict the future or to ascertain the impact of one variable on another.
Prescriptive analytics Techniques that analyze input data and yield a best course of action.
Rule-based model A prescriptive model that is based on a rule or set of rules.
Simulation The use of probability and statistics to construct a computer model to study the impact of uncertainty on the decision at hand.
Simulation optimization The use of probability and statistics to model uncertainty, combined with optimization techniques, to find good decisions in highly complex and highly uncertain settings.
Strategic decision A decision that involves higher-level issues and that is concerned with the overall direction of the organization, defining the overall goals and aspirations for the organization's future.
Supervised learning A category of data-mining techniques in which an algorithm learns how to predict an outcome variable of interest.
Tactical decision A decision concerned with how the organization should achieve the goals and objectives set by its strategy.
Unsupervised learning A category of data-mining techniques in which an algorithm explains relationships without an outcome variable to guide the process.
Utility theory The study of the total worth or relative desirability of a particular outcome that reflects the decision maker's attitude toward a collection of factors such as profit, loss, and risk.

Problems

1. **Quality Versus Low Cost.** Car manufacturer Lexus competes in the luxury car category and is known for its quality. The decision to compete on quality rather than low cost is an example of what type of decision: strategic, tactical, or operational? Explain. LO 1
2. **Package Delivery.** Every day, logistics companies such as United Parcel Service (UPS) must decide how to route their trucks, that is, the order in which to deliver the packages that have been loaded on a truck. UPS developed a system called ORION (On-Road Integrated Optimization and Navigation) which dynamically routes its trucks based on packages on the truck and traffic conditions. It is estimated that ORION saves UPS about 10 million gallons of fuel per year. Is the decision of how to route a truck a strategic, tactical, or operational decision? Explain. LO 1
3. **Airline Decisions.** JetBlue is a major U.S. airline that ranked seventh in the U.S. in 2021 based on number of passengers carried. JetBlue has 270 aircraft and services

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

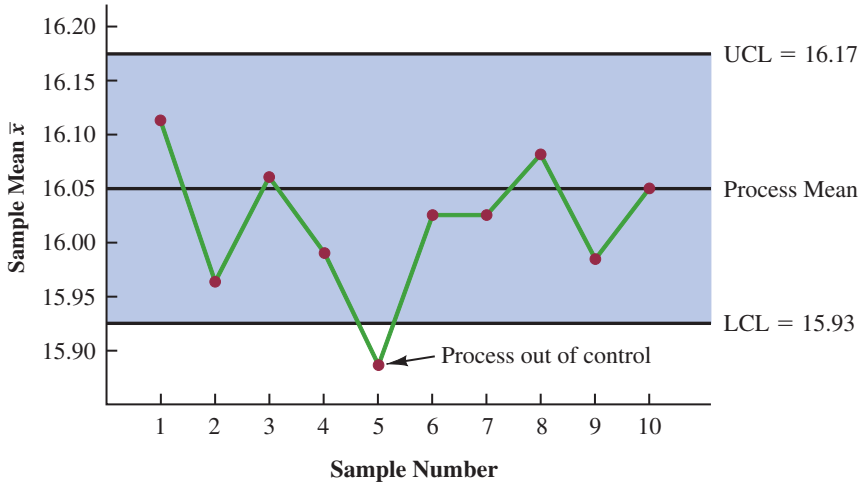
104 destinations. Consider three decisions below that an airline such as JetBlue must make. For each decision, tell if it is strategic, tactical, or operational and explain. LO 1

a. Which crew should be assigned to a given flight?
b. Should we compete on low cost or quality?
c. What type of aircraft should we purchase?

4. **Choosing a College.** Millions of high school graduates decide annually to enroll in a college or university. Many factors can influence a student's choice of where to enroll. List the five steps of the decision-making process and discuss each step of the process, where the decision to be made is choosing which college to attend. LO 2

5. **Package Delivery (Revisited).** Consider again the ORION system discussed in problem 2. Is ORION an example of descriptive, predictive, or prescriptive analytics? Explain. LO 3

6. **Quality Control for Boxes of Cereal.** A control chart is a graphical tool to help determine if a process may be exhibiting non-random variation that needs to be investigated. If non-random variation is present, the process is said to be "out of control" otherwise it is said to be "in control." The following figure shows a control chart for a production line that fills boxes of cereal. Based on past data, we can calculate the mean weight of a box of cereal when the process is in control. The mean weight is 16.05 ounces. We can also calculate control limits, an upper control limit (UCL) and a lower control limit (LCL). New samples are collected over time and the data indicates that the process is in control so long as the new sample weights are between UCL and LCL. As shown in the chart, only sample 5 is outside of the control limits. LO 3



a. Is the control chart an example of descriptive, predictive, or prescriptive analytics?
b. Suppose the control cart is part of a data dashboard and the chart is combined with a rule that does the following. If four consecutive sample mean weights are outside of the control limits, the production line is automatically stopped, and a message appears on the dashboard. The message says "The production line is stopped. The process may be out of control. Please inspect the fill machine." Is this new enhanced control chart combined with a rule an example of descriptive, predictive, or prescriptive analytics?

7. **Amazon Books.** An example of a response from Amazon when this textbook you are reading was chosen online follows. It indicates that some people who purchased this text also tended to purchase *The World Is Flat, Fundamentals of Corporate Finance, and Strategic Marketing Management*. In this application of analytics, is Amazon using descriptive, predictive, or prescriptive analytics? Explain. LO 3

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



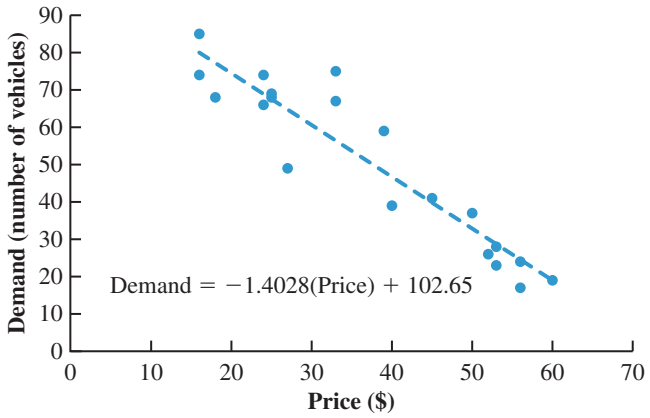
8. **Employee Retention.** Human resource (HR) analytics or people analytics are terms used for the use of analytics to manage a company's workforce. Google, Microsoft, and Walmart, for example, use people analytics to help retain their best people, ensure a diverse workforce, and better understand the strengths and areas needing improvement. **LO 3, 4**
   a. One application of people analytics is to build a model that estimates the probability of an employee departing the company within the next six months. Inputs to the model could be market demand for the skills the person possesses, how long the person has been with the company, and a major life event recently occurring for the person (e.g., divorce). Is this type of model descriptive, predictive, or prescriptive? Explain.
   b. How could you use the model described in part (a) to help improve the workforce?
9. **Supermarket Checkout Lanes.** A supermarket has been experiencing long lines during peak periods of the day. The problem is noticeably worse on certain days of the week, and the peak periods are sometimes different according to the day of the week. There are usually enough workers on the job to open all checkout lanes. The difficulty for store managers is knowing when to call some of the workers stocking shelves up to the front of the store to work the checkout lanes. **LO 4**
   a. How could analytics be used to help the supermarket?
   b. What data would be needed?
10. **Pricing Subcompact Rental Cars.** Setting the right price for a product is an important business decision. If the product is priced too high, the demand could be very low. Set the price too low, demand may be high, but we are potentially leaving money on the table because the revenue per unit is low. Pricing analytics involves finding the right tradeoff between price charged and demand so as to maximize revenue. Suppose we need to set the price for renting a subcompact automobile for one day. Let us outline the decision process:

    Step 1. Identify and define the problem. We need to set a price per day for a midsize rental car.
    Step 2. Determine the criteria that will be used to evaluate alternative solutions. Our goal in setting the price is to maximize revenue per day.
    Step 3. Determine the set of alternative solutions. Based on historical data and the competition, we will consider a broad price range from $10 per day to $60 per day per car.
    Step 4. Evaluate the alternatives. We will evaluate proposed prices based on the expected revenue per day.
    Step 5. Choose an alternative. We will choose the price that maximizes expected revenue.

    We can use data and analytics to complete steps 4 and 5 of the decision process. **LO 3, 4**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

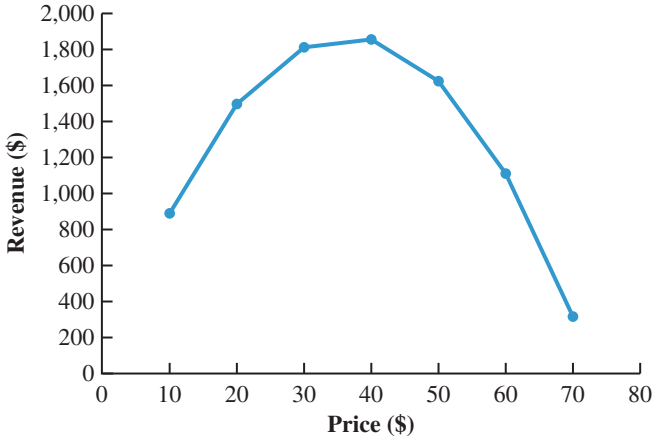
a. Based on historical or test-market data, we can estimate a model that gives expected revenue as a function of price, as shown below. The dots represent the data (price and demand combinations) and the estimated model is the line in the chart: Demand = −1.4028(Price) + 102.65. For example, for price of $35, Demand = −1.4028(35) + 102.65 = 53.552 vehicles. So, we estimate that at a price of $35 per day, the demand will be about 54 vehicles. Is this estimated equation, a descriptive, predictive, or prescriptive model? Explain.



Our goal (step 5) is to find the price that maximizes expected revenue. Revenue = demand × price which is (−1.4028(Price) + 102.65) × (Price) = −1.40228(Price)² + 102.65(Price). The revenue as a function of price is shown below for $10 increments of price.

b. What is the price that maximizes revenue?

c. Is step 5, visually inspecting the revenue function to find a revenue-maximizing price, descriptive, predictive, or prescriptive analytics? Explain.



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 2

Descriptive Statistics

Contents

Analytics in Action:
U.S. Census Bureau

2.1 Overview of Using Data: Definitions and Goals

2.2 Types of Data
- Population and Sample Data
- Quantitative and Categorical Data
- Cross-Sectional and Time Series Data
- Sources of Data

2.3 Exploring Data in Excel
- Sorting and Filtering Data in Excel
- Conditional Formatting of Data in Excel

2.4 Creating Distributions from Data
- Frequency Distributions for Categorical Data
- Relative Frequency and Percent Frequency Distributions
- Frequency Distributions for Quantitative Data
- Histograms
- Frequency Polygons
- Cumulative Distributions

2.5 Measures of Location
- Mean (Arithmetic Mean)
- Median
- Mode
- Geometric Mean

2.6 Measures of Variability
- Range
- Variance
- Standard Deviation
- Coefficient of Variation

2.7 Analyzing Distributions
- Percentiles
- Quartiles
- *z*-Scores
- Empirical Rule
- Identifying Outliers
- Boxplots

2.8 Measures of Association Between Two Variables
- Scatter Charts
- Covariance
- Correlation Coefficient

Summary 68
Glossary 69
Problems 70

Available in Cengage eBook:
- Appendix: Descriptive Statistics with R

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Identify and describe different types of data, including population and sample data, quantitative and categorical data, and cross-sectional and time-series data.

LO 2 Generate insights through sorting, filtering, and conditional formatting data.

LO 3 Construct and interpret frequency, relative frequency, and percent frequency distributions for categorical data.

LO 4 Construct and interpret frequency, relative frequency, and percent frequency distributions for quantitative data.

LO 5 Construct and interpret histograms and frequency polygons to visualize the distribution of quantitative data.

LO 6 Construct and interpret cumulative frequency, cumulative relative frequency, and cumulative percent frequency distributions for quantitative data.

LO 7 Interpret the shape of a distribution of data and identify positive skewness, negative skewness, and symmetric distributions.

LO 8 Calculate and interpret measures of location such as the mean, median, geometric mean, and mode.

LO 9 Calculate and interpret measures of variability such as the range, variance, standard deviation, and coefficient of variation.

LO 10 Analyze and interpret distributions of data using percentiles, quartiles, *z*-scores, and the empirical rule.

LO 11 Identify outliers in a set of data.

LO 12 Construct and interpret a boxplot.

LO 13 Create and interpret a scatter chart for two quantitative variables.

LO 14 Calculate and interpret the covariance and correlation coefficient for two quantitative variables.

Analytics in Action

U.S. Census Bureau

The U.S. Census Bureau is part of the Department of Commerce. The U.S. Census Bureau collects data related to the population and economy of the United States using a variety of methods and for many purposes. These data are essential to many government and business decisions.

Probably the best-known data collected by the U.S. Census Bureau is the decennial census, which is an effort to count the total U.S. population. Collecting these data is a huge undertaking involving mailings, door-to-door visits, and other methods. The decennial census collects categorical data such as the sex and race of the respondents, as well as quantitative data such as the number of people living in the household. The data collected in the decennial census are used to determine the number of representatives assigned to each state, the number of Electoral College votes apportioned to each state, and how federal government funding is divided among communities.

The U.S. Census Bureau also administers the Current Population Survey (CPS). The CPS is a cross-sectional monthly survey of a sample of 60,000 households used to estimate employment and unemployment rates in different geographic areas. The CPS has been administered since 1940, so an extensive time series of employment and unemployment data now exists. These data drive government policies such as job assistance programs. The estimated unemployment rates are watched closely as an overall indicator of the health of the U.S. economy.

The data collected by the U.S. Census Bureau are also very useful to businesses. Retailers use data on population changes in different areas to plan new store openings. Mail-order catalog companies use the demographic data when designing targeted marketing campaigns. In many cases, businesses combine the data collected by the U.S. Census Bureau with their own data on customer behavior to plan strategies and to identify potential customers. The data collected by the U.S. Census Bureau is publicly available and can be downloaded from its web site.

In this chapter, we first explain the need to collect and analyze data and identify some common sources of data. Then we discuss the types of data that you may encounter in practice and present several numerical measures for summarizing data. We cover some common ways of manipulating and summarizing data using spreadsheets. We then develop numerical summary measures for data sets consisting of a single variable. When a data set contains more than one variable, the same numerical measures can be computed separately for each variable. In the two-variable case, we also develop measures of the relationship between the variables.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2.1 Overview of Using Data: Definitions and Goals

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation. Table 2.1 shows a data set containing information for stocks in the Dow Jones Industrial Index (or simply "the Dow") on June 3, 2022. The Dow is tracked by many financial advisors and investors as an indication of the state of the overall financial markets and the economy in the United States. The share prices for the 30 companies listed in Table 2.1 are the basis for computing the Dow Jones Industrial Average (DJI), which is tracked continuously by virtually every financial publication. The index is named for Charles Dow and Edward Jones who first began calculating the DJI in 1896.

A characteristic or a quantity of interest that can take on different values is known as a **variable**; for the data in Table 2.1, the variables are Symbol, Industry Sector, and Share Price. An **observation** is a set of values corresponding to a set of variables; each row in Table 2.1 corresponds to an observation.

Table 2.1 Data for Dow Jones Industrial Index Companies

Company	Symbol	Industry Sector	Share Price ($)
3M Co.	MMM	Conglomerate	145.88
American Express Co.	AXP	Financial Services	166.80
Amgen Inc.	AMGN	Healthcare	248.45
Apple Inc.	AAPL	Technology	145.38
Boeing Co.	BA	Industrials	139.17
Caterpillar Inc.	CAT	Industrials	222.96
Chevron Corp.	CVX	Energy	177.66
Cisco Systems Inc.	CSCO	Technology	45.25
Coca-Cola Co.	KO	Consumer Defensive	62.99
Dow Inc.	DOW	Basic Materials	67.05
Goldman Sachs Group Inc.	GS	Financial Services	318.64
Home Depot Inc.	HD	Consumer Cyclical	305.00
Honeywell International Inc.	HON	Industrials	194.82
Intel Corp.	INTC	Technology	43.39
International Business Machines Corp.	IBM	Technology	141.11
Johnson & Johnson	JNJ	Healthcare	176.43
JPMorgan Chase & Co.	JPM	Financial Services	129.99
McDonald's Corp.	MCD	Consumer Cyclical	248.19
Merck & Co. Inc.	MRK	Healthcare	89.89
Microsoft Corp.	MSFT	Technology	270.02
Nike Inc. Cl B	NKE	Consumer Cyclical	120.98
Procter & Gamble Co.	PG	Consumer Defensive	145.97
Salesforce Inc.	CRM	Technology	184.93
Travelers Cos. Inc.	TRV	Financial Services	174.05
UnitedHealth Group Inc.	UNH	Healthcare	485.44
Verizon Communications Inc.	VZ	Communication Services	50.78
Visa Inc.	V	Financial Services	212.61
Walgreens Boots Alliance Inc.	WBA	Healthcare	43.14
Walmart Inc.	WMT	Consumer Defensive	125.33
Walt Disney Co.	DIS	Communication Services	108.69

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Decision variables used in optimization models are covered in Chapters 14, 15, and 16. Random variables are covered in greater detail in Chapters 5 and 13.

Practically every problem (and opportunity) that an organization (or individual) faces is concerned with the impact of the possible values of relevant variables on the business outcome. Thus, we are concerned with how the value of a variable can vary; **variation** is the difference in a variable measured over observations (time, customers, items, etc.).

The role of descriptive analytics is to collect and analyze data to gain a better understanding of variation and its impact on the business setting. The values of some variables are under direct control of the decision maker (these are often called decision variables). The values of other variables may fluctuate with uncertainty because of factors outside the direct control of the decision maker. In general, a quantity whose values are not known with certainty is called a **random variable, or uncertain variable**. When we collect data, we are gathering past observed values, or realizations of a variable. By collecting these past realizations of one or more variables, our goal is to learn more about the variation of a particular business situation.

2.2 Types of Data

Population and Sample Data

To ensure that the companies in the Dow form a representative sample, companies are periodically added and removed from the Dow. It is possible that the companies in the Dow today have changed from what is shown in Table 2.1.

Data can be categorized in several ways based on how they are collected and the type collected. In many cases, it is not feasible to collect data from the **population** of all elements of interest. In such instances, we collect data from a subset of the population known as a **sample**. For example, with the thousands of publicly traded companies in the United States, tracking and analyzing all of these stocks every day would be too time consuming and expensive. The Dow represents a sample of 30 stocks of large public companies based in the United States, and it is often interpreted to represent the larger population of all publicly traded companies. It is very important to collect sample data that are representative of the population data so that generalizations can be made from them. In most cases (although not true of the Dow), a representative sample can be gathered by **random sampling** from the population data. Dealing with populations and samples can introduce subtle differences in how we calculate and interpret summary statistics. In almost all practical applications of business analytics, we will be dealing with sample data.

Quantitative and Categorical Data

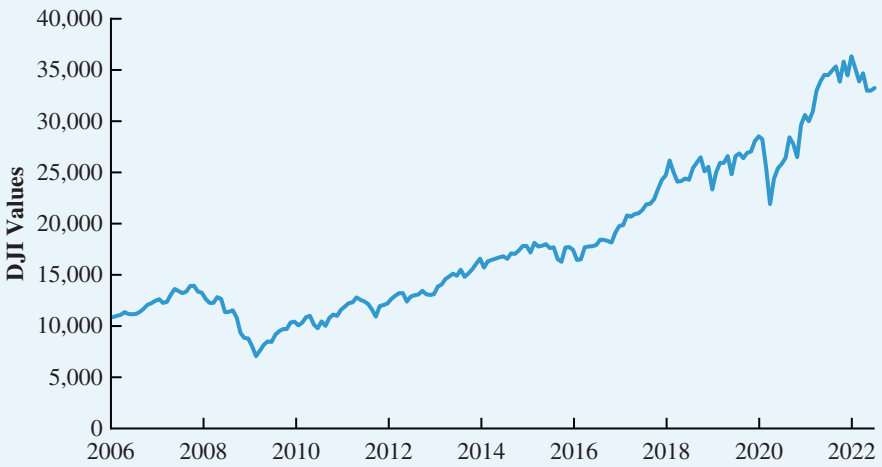
Data are considered **quantitative data** if numeric and arithmetic operations, such as addition, subtraction, multiplication, and division, can be performed on them. For instance, we can sum the values for Share Price ($) and then divide by the number of stocks in the DJI from the Dow data in Table 2.1 to calculate an average stock price of all shares traded by companies included in the Dow. If arithmetic operations cannot be performed on the data, they are considered **categorical data**. We can summarize categorical data by counting the number of observations or computing the proportions of observations in each category. For instance, the data in the Industry Sector column in Table 2.1 are categorical. We can count the number of companies in the Dow that are in specific industry sectors. Table 2.1 shows five companies in the financial services industry sector: American Express, Goldman Sachs Group, Inc., JPMorgan Chase & Co., Travelers Cos., Inc., and Visa, Inc. We cannot perform arithmetic operations on the data in the Industry Sector column.

Cross-Sectional and Time Series Data

For statistical analysis, it is important to distinguish between cross-sectional data and time series data. **Cross-sectional data** are collected from several entities at the same, or approximately the same, point in time. The data in Table 2.1 are cross-sectional because they describe the 30 companies that comprise the Dow at the same point in time (June 3, 2022). **Time series data** are collected over several time periods. Graphs of time series data are frequently found in business and economic publications. Such graphs help analysts understand what happened in the past, identify trends over time, and project future levels for the time series. For example, the graph of the time series in Figure 2.1 shows the DJI value from January 2006 to June 2022. Figure 2.1 shows that the DJI has been

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.1 Dow Jones Industrial Average Values Since 2006



generally increasing since 2006 with a few periods of decline. From 2007 to 2009 the DJI declined in value by nearly 50% due to the financial crisis in 2008. Additional declines are seen in 2019 to 2021 due to the onset of the COVID-19 pandemic and a shallower decline in 2022 due to lingering economic effects of the pandemic combined with other economic factors.

Sources of Data

Data necessary to analyze a business problem or opportunity can often be obtained with an appropriate study; such statistical studies can be classified as either experimental or observational. In an *experimental study*, a variable of interest is first identified. Then one or more other variables are identified and controlled or manipulated to obtain data about how these variables influence the variable of interest. For example, if a pharmaceutical firm conducts an experiment to learn about how a new drug affects blood pressure, then blood pressure is the variable of interest. The dosage level of the new drug is another variable that is hoped to have a causal effect on blood pressure. To obtain data about the effect of the new drug, researchers select a sample of individuals. The dosage level of the new drug is controlled by giving different dosages to the different groups of individuals. Before and after the study, data on blood pressure are collected for each group. Statistical analysis of these experimental data can help determine how the new drug affects blood pressure.

Nonexperimental, or *observational*, *studies* make no attempt to control the variables of interest. A survey is perhaps the most common type of observational study. For instance, in a personal interview survey, research questions are first identified. Then a questionnaire is designed and administered to a sample of individuals. Some restaurants use observational studies to obtain data about customer opinions with regard to the quality of food, quality of service, atmosphere, and so on. A customer opinion questionnaire used by Chops City Grill in Naples, Florida, is shown in Figure 2.2. Note that the customers who fill out the questionnaire are asked to provide ratings for 12 variables, including overall experience, the greeting by hostess, the table visit by the manager, overall service, and so on. The response categories of excellent, good, average, fair, and poor provide categorical data that enable Chops City Grill management to maintain high standards for the restaurant's food and service.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.2 Customer Opinion Questionnaire Used by Chops City Grill Restaurant

Chops CITY GRILL

Date: _____ Server Name: _____

Our customers are our top priority. Please take a moment to fill out our survey card, so we can better serve your needs. You may return this card to the front desk or return by mail. Thank you!

SERVICE SURVEY	Excellent	Good	Average	Fair	Poor
Overall Experience	❑	❑	❑	❑	❑
Greeting by Hostess	❑	❑	❑	❑	❑
Manager (Table Visit)	❑	❑	❑	❑	❑
Overall Service	❑	❑	❑	❑	❑
Professionalism	❑	❑	❑	❑	❑
Menu Knowledge	❑	❑	❑	❑	❑
Friendliness	❑	❑	❑	❑	❑
Wine Selection	❑	❑	❑	❑	❑
Menu Selection	❑	❑	❑	❑	❑
Food Quality	❑	❑	❑	❑	❑
Food Presentation	❑	❑	❑	❑	❑
Value for $ Spent	❑	❑	❑	❑	❑

What comments could you give us to improve our restaurant?

_____

_____

Thank you, we appreciate your comments. —The staff of Chops City Grill.

In some cases, the data needed for a particular application exist from an experimental or observational study that has already been conducted. For example, companies maintain a variety of databases about their employees, customers, and business operations. Data on employee salaries, ages, and years of experience can usually be obtained from internal personnel records. Other internal records contain data on sales, advertising expenditures, distribution costs, inventory levels, and production quantities. Most companies also maintain detailed data about their customers.

In Chapter 17, we discuss methods for determining the value of additional information that can be provided by collecting data.

Anyone who wants to use data and statistical analysis to aid in decision making must be aware of the time and cost required to obtain the data. The use of existing data sources is desirable when data must be obtained in a relatively short period of time. If important data are not readily available from a reliable existing source, the additional time and cost involved in obtaining the data must be taken into account. In all cases, the decision maker should consider the potential contribution of the statistical analysis to the decision-making process. The cost of data acquisition and the subsequent statistical analysis should not exceed the savings generated by using the information to make a better decision.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notes + Comments

1. Organizations that specialize in collecting and maintaining data make available substantial amounts of business and economic data. Companies can access these external data sources through leasing arrangements or by purchase. Dun & Bradstreet, Bloomberg, and Dow Jones & Company are three firms that provide extensive business database services to clients. Nielsen and Ipsos are two companies that have built successful businesses collecting and processing data that they sell to advertisers and product manufacturers. Data are also available from a variety of industry associations and special-interest organizations.
2. Government agencies are another important source of existing data. For instance, the web site *data.gov* was launched by the U.S. government in 2009 to make it easier for the public to access data collected by the U.S. federal government. The *data.gov* web site includes over 150,000 data sets from a variety of U.S. federal departments and agencies, but many other federal agencies maintain their own web sites and data repositories. Many state and local governments are also now providing data sets online. As examples, the states of California and Texas maintain open data portals at *data.ca.gov* and *data.texas.gov*, respectively. New York City's open data web site is *opendata.cityofnewyork.us* and the city of Cincinnati, Ohio, is at *data.cincinnati-oh.gov*. In general, the Internet is an important source of data and statistical information. One can obtain access to stock quotes, meal prices at restaurants, salary data, and a wide array of other information simply by performing an Internet search.
3. Google provides a data set search tool at *datasetsearch.research.google.com*. This website allows the user to search for data sets based on keywords.

2.3 Exploring Data in Excel

Projects often involve so much data that it is difficult to analyze all of the data at once. In this section, we examine methods for exploring data using Excel to develop insights.

Sorting and Filtering Data in Excel

Excel contains many useful features for sorting and filtering data so that one can more easily identify patterns. Table 2.2 contains data on the 20 top-selling vehicles in the United States in based on 2021 sales. The table shows the model and manufacturer of each vehicle as well as the sales for the model in 2021 and 2020.

Figure 2.3 shows the data from Table 2.2 entered into an Excel spreadsheet, and the percent change in sales for each model from 2020 to 2021 has been calculated. This is done by entering the formula *=(D2-E2)/E2* in cell F2 and then copying the contents of this cell to cells F3 to F20.

Suppose that we want to sort these vehicles by 2020 sales instead of by 2021 sales. To do this, we use Excel's Sort function, as shown in the following steps.

Step 1. Select cells A1:F21
Step 2. Click the **Data** tab in the Ribbon
Step 3. Click **Sort** in the **Sort & Filter** group
Step 4. Select the check box for **My data has headers**
Step 5. In the first **Sort by** dropdown menu, select **Sales (2020)**
Step 6. In the **Order** dropdown menu, select **Largest to Smallest** (see Figure 2.4)
Step 7. Click **OK**

The result of using Excel's Sort function for the 2020 sales data is shown in Figure 2.5. From Figures 2.3 and 2.5, we see that, the Ford F-Series vehicle was the top selling vehicle in 2020 and 2021, and the Chevrolet Silverado outsold the Ram Pickup in 2020 but not in 2021. Note that while we sorted on **Sales (2020)**, which is in column E, the data in all other columns are adjusted accordingly.

Now let's suppose that we are interested only in seeing the sales of models made by Toyota. We can do this using Excel's Filter function:

Step 1. Select cells A1:F21
Step 2. Click the **Data** tab in the Ribbon

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 2.2 20 Top-Selling Vehicles in the United States Based on 2021 Sales

Rank (by 2021 Sales)	Manufacturer	Model	Sales (2021)	Sales (2020)
1	Ford	F-Series	726,004	787,423
2	Ram	Pickup	569,388	563,750
3	Chevrolet	Silverado	519,774	586,652
4	Toyota	RAV4	407,739	430,559
5	Honda	CR-V	361,271	333,584
6	Toyota	Camry	313,795	294,367
7	Nissan	Rogue	285,602	227,935
8	Jeep	Grand Cherokee	264,444	209,710
9	Toyota	Highlander	264,128	212,322
10	Honda	Civic	263,787	261,175
11	Toyota	Tacoma	252,520	238,903
12	GMC	Sierra	248,924	252,972
13	Toyota	Corolla	229,785	217,806
14	Ford	Explorer	219,871	226,205
15	Jeep	Wrangler	204,609	201,387
16	Honda	Accord	202,676	199,484
17	Tesla	Model Y	190,395	65,405
18	Mazda	CX-5	168,448	146,477
19	Chevrolet	Equinox	165,323	271,021
20	Subaru	Forester	154,723	177,029

DATA*file*
top20vehicles2021

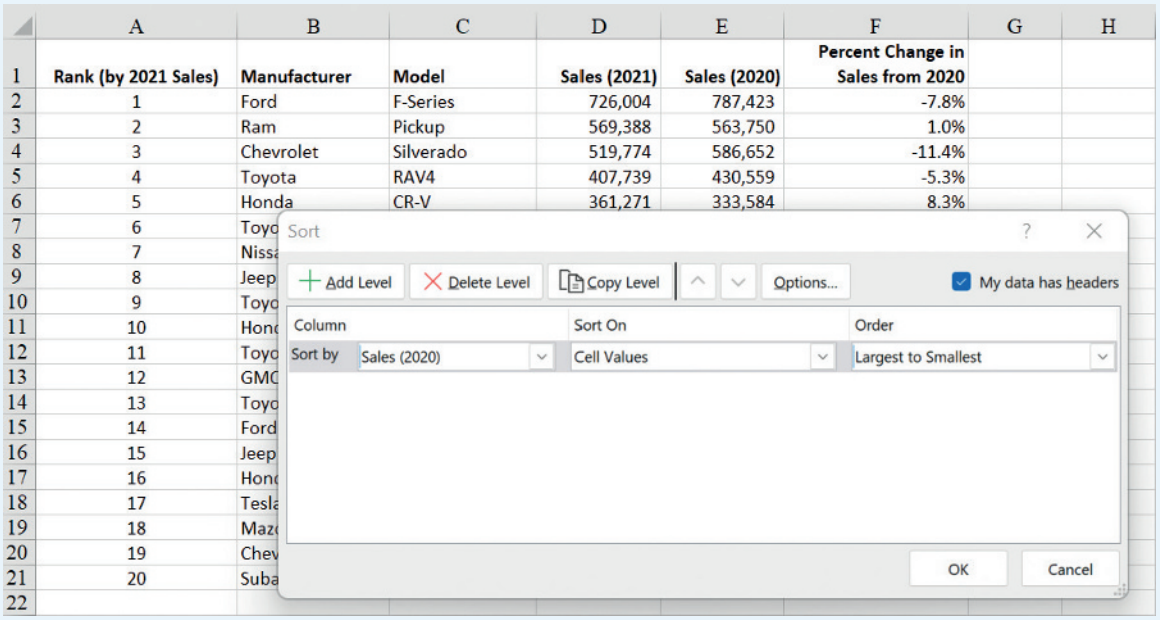
Source: *Forbes* website

Figure 2.3 Data for 20 Top-Selling Vehicles Entered into Excel with Percent Change in Sales from 2020 to 2021

	A	B	C	D	E	F
1	Rank (by 2021 Sales)	Manufacturer	Model	Sales (2021)	Sales (2020)	Percent Change in Sales from 2020
2	1	Ford	F-Series	726,004	787,423	-7.8%
3	2	Ram	Pickup	569,388	563,750	1.0%
4	3	Chevrolet	Silverado	519,774	586,652	-11.4%
5	4	Toyota	RAV4	407,739	430,559	-5.3%
6	5	Honda	CR-V	361,271	333,584	8.3%
7	6	Toyota	Camry	313,795	294,367	6.6%
8	7	Nissan	Rogue	285,602	227,935	25.3%
9	8	Jeep	Grand Cherokee	264,444	209,710	26.1%
10	9	Toyota	Highlander	264,128	212,322	24.4%
11	10	Honda	Civic	263,787	261,175	1.0%
12	11	Toyota	Tacoma	252,520	238,903	5.7%
13	12	GMC	Sierra	248,924	252,972	-1.6%
14	13	Toyota	Corolla	229,785	217,806	5.5%
15	14	Ford	Explorer	219,871	226,205	-2.8%
16	15	Jeep	Wrangler	204,609	201,387	1.6%
17	16	Honda	Accord	202,676	199,484	1.6%
18	17	Tesla	Model Y	190,395	65,405	191.1%
19	18	Mazda	CX-5	168,448	146,477	15.0%
20	19	Chevrolet	Equinox	165,323	271,021	-39.0%
21	20	Subaru	Forester	154,723	177,029	-12.6%

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.4 Using Excel's Sort Function to Sort the Top-Selling Vehicles Data



	A	B	C	D	E	F	G	H
1	Rank (by 2021 Sales)	Manufacturer	Model	Sales (2021)	Sales (2020)	Percent Change in Sales from 2020		
2	1	Ford	F-Series	726,004	787,423	-7.8%		
3	2	Ram	Pickup	569,388	563,750	1.0%		
4	3	Chevrolet	Silverado	519,774	586,652	-11.4%		
5	4	Toyota	RAV4	407,739	430,559	-5.3%		
6	5	Honda	CR-V	361,271	333,584	8.3%		
7	6	Toyo						
8	7	Nissa						
9	8	Jeep						
10	9	Toyo						
11	10	Hond						
12	11	Toyo						
13	12	GMC						
14	13	Toyo						
15	14	Ford						
16	15	Jeep						
17	16	Hond						
18	17	Tesla						
19	18	Mazd						
20	19	Chev						
21	20	Suba						
22								

Figure 2.5 Top-Selling Vehicles Data Sorted by 2020 Sales

	A	B	C	D	E	F
1	Rank (by 2021 Sales)	Manufacturer	Model	Sales (2021)	Sales (2020)	Percent Change in Sales from 2020
2	1	Ford	F-Series	726,004	787,423	-7.8%
3	3	Chevrolet	Silverado	519,774	586,652	-11.4%
4	2	Ram	Pickup	569,388	563,750	1.0%
5	4	Toyota	RAV4	407,739	430,559	-5.3%
6	5	Honda	CR-V	361,271	333,584	8.3%
7	6	Toyota	Camry	313,795	294,367	6.6%
8	19	Chevrolet	Equinox	165,323	271,021	-39.0%
9	10	Honda	Civic	263,787	261,175	1.0%
10	12	GMC	Sierra	248,924	252,972	-1.6%
11	11	Toyota	Tacoma	252,520	238,903	5.7%
12	7	Nissan	Rogue	285,602	227,935	25.3%
13	14	Ford	Explorer	219,871	226,205	-2.8%
14	13	Toyota	Corolla	229,785	217,806	5.5%
15	9	Toyota	Highlander	264,128	212,322	24.4%
16	8	Jeep	Grand Cherokee	264,444	209,710	26.1%
17	15	Jeep	Wrangler	204,609	201,387	1.6%
18	16	Honda	Accord	202,676	199,484	1.6%
19	20	Subaru	Forester	154,723	177,029	-12.6%
20	18	Mazda	CX-5	168,448	146,477	15.0%
21	17	Tesla	Model Y	190,395	65,405	191.1%

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.6 Top-Selling Vehicles Data Filtered to Show Only Vehicles Manufactured by Toyota

	A	B	C	D	E	F
1	Rank (by 2021 Sales)	Manufacturer	Model	Sales (2021)	Sales (2020)	Percent Change in Sales from 2020
5	4	Toyota	RAV4	407,739	430,559	-5.3%
7	6	Toyota	Camry	313,795	294,367	6.6%
10	9	Toyota	Highlander	264,128	212,322	24.4%
12	11	Toyota	Tacoma	252,520	238,903	5.7%
14	13	Toyota	Corolla	229,785	217,806	5.5%
22						

Step 3. Click **Filter** in the **Sort & Filter** group
Step 4. Click on the **Filter Arrow** in column B, next to **Manufacturer**
Step 5. If all choices are checked, you can easily deselect all choices by unchecking (**Select All**). Then select only the check box for **Toyota**.
Step 6. Click **OK**

The result is a display of only the data for vehicles made by Toyota (see Figure 2.6). We now see that of the 20 top-selling vehicles in 2021, Toyota made five of them: the RAV4, the Camry, the Highlander, the Tacoma, and the Corolla. Note that when you Filter the data, Excel hides the rows for values that have been filtered out of the display. We can further filter the data by choosing the down arrows in the other columns. We can make all data visible again by clicking on the down arrow in column B and checking **(Select All)** and clicking **OK**, or by clicking **Filter** in the **Sort & Filter** Group again from the **Data** tab.

Conditional Formatting of Data in Excel

Conditional formatting in Excel can make it easy to identify data that satisfy certain conditions in a data set. For instance, suppose that we wanted to quickly identify the vehicles in Table 2.2 for which sales had decreased from 2020 to 2021. We can quickly highlight these vehicles by using the steps that follow.

Step 1. Starting with the original data shown in Figure 2.3, select cells F1:F21
Step 2. Click the **Home** tab in the Ribbon
Step 3. Click **Conditional Formatting** in the **Styles** group
Step 4. Select **Highlight Cells Rules**, and click **Less Than . . .** from the dropdown menu
Step 5. Enter *0%* in the **Format cells that are LESS THAN:** box and **Light Red Fill with Dark Red** Text in the **with** box
Step 6. Click **OK**

The results are shown in Figure 2.7. Here we see that the vehicles with decreasing sales (Ford F-Series, Chevrolet Silverado, Toyota RAV4, GMC Sierra, Ford Explorer, Chevrolet Equinox, and Subaru Forester) are now clearly visible. Although not shown here, it is interesting to note that each of these vehicles with declining sales from 2020 to 2021 are classified as either trucks or sports-utility vehicles (SUVs). This demonstrates how data exploration using conditional formatting can help derive insights from data.

Bar charts and other graphical presentations are covered in detail in Chapter 3.

Excel's Conditional Formatting function offers tremendous flexibility. Instead of highlighting only vehicles with decreasing sales, we could instead choose **Data Bars** from the **Conditional Formatting** dropdown menu in the **Styles** Group of the **Home** tab in the Ribbon. The result of using the **Blue Data Bar Gradient Fill** option is shown in Figure 2.8. Data bars are essentially a bar chart input into the cells that shows the magnitude of the cell values. The widths of the bars in this display are comparable to the values of the variable for which the bars have been drawn; a value of 20 creates a bar twice as wide as that for a value of 10. Negative values are shown to the left side of the axis; positive values are shown to the right. Cells with negative values are shaded in red, and those with positive values are shaded

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.7 Using Conditional Formatting in Excel to Highlight Vehicles with Declining Sales from 2020 to 2021

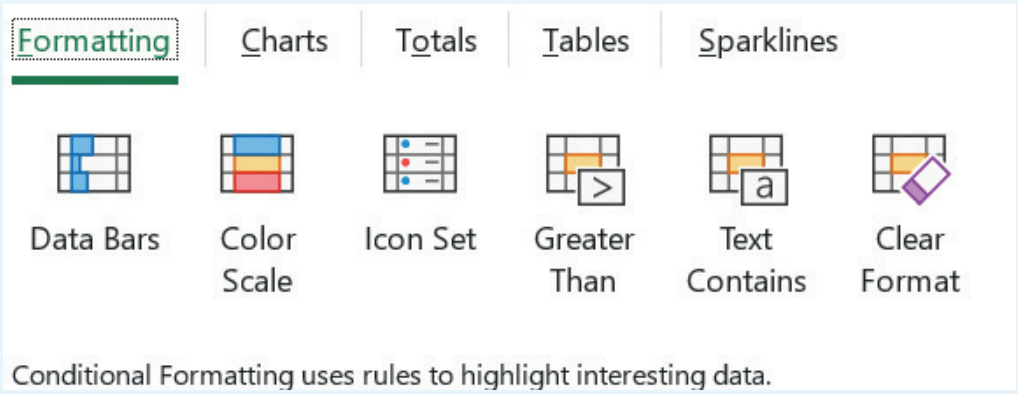
	A	B	C	D	E	F
1	Rank (by 2021 Sales)	Manufacturer	Model	Sales (2021)	Sales (2020)	Percent Change in Sales from 2020
2	1	Ford	F-Series	726,004	787,423	-7.8%
3	2	Ram	Pickup	569,388	563,750	1.0%
4	3	Chevrolet	Silverado	519,774	586,652	-11.4%
5	4	Toyota	RAV4	407,739	430,559	-5.3%
6	5	Honda	CR-V	361,271	333,584	8.3%
7	6	Toyota	Camry	313,795	294,367	6.6%
8	7	Nissan	Rogue	285,602	227,935	25.3%
9	8	Jeep	Grand Cherokee	264,444	209,710	26.1%
10	9	Toyota	Highlander	264,128	212,322	24.4%
11	10	Honda	Civic	263,787	261,175	1.0%
12	11	Toyota	Tacoma	252,520	238,903	5.7%
13	12	GMC	Sierra	248,924	252,972	-1.6%
14	13	Toyota	Corolla	229,785	217,806	5.5%
15	14	Ford	Explorer	219,871	226,205	-2.8%
16	15	Jeep	Wrangler	204,609	201,387	1.6%
17	16	Honda	Accord	202,676	199,484	1.6%
18	17	Tesla	Model Y	190,395	65,405	191.1%
19	18	Mazda	CX-5	168,448	146,477	15.0%
20	19	Chevrolet	Equinox	165,323	271,021	-39.0%
21	20	Subaru	Forester	154,723	177,029	-12.6%

Figure 2.8 Using Conditional Formatting in Excel to Generate Data Bars for the Top-Selling Vehicles Data

	A	B	C	D	E	F
1	Rank (by 2021 Sales)	Manufacturer	Model	Sales (2021)	Sales (2020)	Percent Change in Sales from 2020
2	1	Ford	F-Series	726,004	787,423	-7.8%
3	2	Ram	Pickup	569,388	563,750	1.0%
4	3	Chevrolet	Silverado	519,774	586,652	-11.4%
5	4	Toyota	RAV4	407,739	430,559	-5.3%
6	5	Honda	CR-V	361,271	333,584	8.3%
7	6	Toyota	Camry	313,795	294,367	6.6%
8	7	Nissan	Rogue	285,602	227,935	25.3%
9	8	Jeep	Grand Cherokee	264,444	209,710	26.1%
10	9	Toyota	Highlander	264,128	212,322	24.4%
11	10	Honda	Civic	263,787	261,175	1.0%
12	11	Toyota	Tacoma	252,520	238,903	5.7%
13	12	GMC	Sierra	248,924	252,972	-1.6%
14	13	Toyota	Corolla	229,785	217,806	5.5%
15	14	Ford	Explorer	219,871	226,205	-2.8%
16	15	Jeep	Wrangler	204,609	201,387	1.6%
17	16	Honda	Accord	202,676	199,484	1.6%
18	17	Tesla	Model Y	190,395	65,405	191.1%
19	18	Mazda	CX-5	168,448	146,477	15.0%
20	19	Chevrolet	Equinox	165,323	271,021	-39.0%
21	20	Subaru	Forester	154,723	177,029	-12.6%

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.9 Excel Quick Analysis Button Formatting Options



in blue. Again, we can easily see which models had decreasing sales, but Data Bars also provide us with a visual representation of the magnitude of the change in sales. Many other Conditional Formatting options are available in Excel.

The **Quick Analysis** button in Excel appears just outside the bottom-right corner of a group of selected cells whenever you select multiple cells. Clicking the **Quick Analysis** button gives you shortcuts for Conditional Formatting, adding Data Bars, and other operations. Clicking on this button gives you the options shown in Figure 2.9 for **Formatting**. Note that there are also tabs for **Charts**, **Totals**, **Tables**, and **Sparklines**.

2.4 Creating Distributions from Data

Distributions help summarize many characteristics of a data set by describing how often certain values for a variable appear in that data set. Distributions can be created for both categorical and quantitative data, and they assist the analyst in gauging variation.

Frequency Distributions for Categorical Data

Bins for categorical data are also referred to as classes.

It is often useful to create a frequency distribution for a data set. A **frequency distribution** is a summary of data that shows the number (frequency) of observations in each of several nonoverlapping classes, typically referred to as **bins**. Consider the data in Table 2.3, taken from a sample of 50 soft drink purchases. Each purchase is for one of five popular soft drinks, which define the five bins: Coca-Cola, Diet Coke, Dr. Pepper, Pepsi, and Sprite.

To develop a frequency distribution for these data, we count the number of times each soft drink appears in Table 2.3. Coca-Cola appears 19 times, Diet Coke appears 8 times, Dr. Pepper appears 5 times, Pepsi appears 13 times, and Sprite appears 5 times. These counts are summarized in the frequency distribution in Table 2.4. This frequency distribution provides a summary of how the 50 soft drink purchases are distributed across the 5 soft drinks. This summary offers more insight than the original data shown in Table 2.3. The frequency distribution shows that Coca-Cola is the leader, Pepsi is second, Diet Coke is third, and Sprite and Dr. Pepper are tied for fourth. The frequency distribution thus summarizes information about the popularity of the five soft drinks.

Absolute and relative references are discussed in Appendix A.

We can use Excel to calculate the frequency of categorical observations occurring in a data set using the COUNTIF function. Figure 2.10 shows the sample of 50 soft drink purchases in an Excel spreadsheet. Column D contains the five different soft drink categories as the bins. In cell E2, we enter the formula =*COUNTIF($A$2:$B$26, D2)*, where A2:B26 is the range for the sample data, and D2 is the bin (Coca-Cola) that we are trying to match. The COUNTIF function in Excel counts the number of times a certain value appears in the indicated range. In this case we want to count the number of times Coca-Cola appears in the sample data. The

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 2.3 Data from a Sample of 50 Soft Drink Purchases

Coca-Cola	Sprite	Pepsi
Diet Coke	Coca-Cola	Coca-Cola
Pepsi	Diet Coke	Coca-Cola
Diet Coke	Coca-Cola	Coca-Cola
Coca-Cola	Diet Coke	Pepsi
Coca-Cola	Coca-Cola	Dr. Pepper
Dr. Pepper	Sprite	Coca-Cola
Diet Coke	Pepsi	Diet Coke
Pepsi	Coca-Cola	Pepsi
Pepsi	Coca-Cola	Pepsi
Coca-Cola	Coca-Cola	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coca-Cola	Coca-Cola
Coca-Cola	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coca-Cola	Pepsi	Sprite
Coca-Cola	Diet Coke	

DATA*file* softdrinks

result is a value of 19 in cell E2, indicating that Coca-Cola appears 19 times in the sample data. We can copy the formula from cell E2 to cells E3 to E6 to get frequency counts for Diet Coke, Pepsi, Dr. Pepper, and Sprite. By using the absolute reference $A$2:$B$26 in our formula, Excel always searches the same sample data for the values we want when we copy the formula.

Relative Frequency and Percent Frequency Distributions

A frequency distribution shows the number (frequency) of items in each of several nonoverlapping bins. However, we are often interested in the proportion, or percentage, of items in each bin. The *relative frequency* of a bin equals the fraction or proportion of items belonging to a class. For a data set with n observations, the relative frequency of each bin can be determined as follows:

The percent frequency of a bin is the relative frequency multiplied by 100.

$$\text{Relative frequency of a bin} = \frac{\text{Frequency of the bin}}{n}$$

A **relative frequency distribution** is a tabular summary of data showing the relative frequency for each bin. A **percent frequency distribution** summarizes the percent frequency of the data for each bin. Table 2.5 shows a relative frequency distribution and a percent frequency distribution for the soft drink data. Using the data from Table 2.4, we see that the relative frequency for Coca-Cola is 19/50 = 0.38, the relative frequency for Diet

Table 2.4 Frequency Distribution of Soft Drink Purchases

Soft Drink	Frequency
Coca-Cola	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.10 Creating a Frequency Distribution for Soft Drinks Data in Excel

	A	B	C	D	E
1	Sample Data			Bins	
2	Coca-Cola	Coca-Cola		Coca-Cola	18
3	Diet Coke	Sprite		Diet Coke	8
4	Pepsi	Pepsi		Dr. Pepper	5
5	Diet Coke	Pepsi		Pepsi	14
6	Coca-Cola	Pepsi		Sprite	5
7	Coca-Cola	Sprite			
8	Dr. Pepper	Dr. Pepper			
9	Diet Coke	Pepsi			
10	Pepsi	Diet Coke			
11	Pepsi	Pepsi			
12	Coca-Cola	Coca-Cola			
13	Dr. Pepper	Coca-Cola			
14	Sprite	Diet Coke			
15	Coca-Cola	Pepsi			
16	Diet Coke	Pepsi			
17	Coca-Cola	Pepsi			
18	Coca-Cola	Coca-Cola			
19	Diet Coke	Dr. Pepper			
20	Coca-Cola	Sprite			
21	Coca-Cola	Coca-Cola			
22	Coca-Cola	Coca-Cola			
23	Sprite	Pepsi			
24	Coca-Cola	Dr. Pepper			
25	Coca-Cola	Pepsi			
26	Diet Coke	Pepsi			

Coke is 8/50 = 0.16, and so on. From the percent frequency distribution, we see that 38% of the purchases were Coca-Cola, 16% were Diet Coke, and so on. We can also note that 38% + 26% + 16% = 80% of the purchases were the top three soft drinks.

A percent frequency distribution can be used to provide estimates of the relative likelihoods of different values for a random variable. So, by constructing a percent frequency distribution from observations of a random variable, we can estimate the probability distribution that characterizes its variability. For example, the volume of soft drinks sold by a concession stand at an upcoming concert may not be known with certainty. However, if the data used to construct Table 2.5 are representative of the concession stand's customer population, then the concession stand manager can use this information to determine the appropriate volume of each type of soft drink.

Table 2.5 Relative Frequency and Percent Frequency Distributions of Soft Drink Purchases

Soft Drink	Relative Frequency	Percent Frequency (%)
Coca-Cola	0.38	38
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi	0.26	26
Sprite	0.10	10
Total	1.00	100

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
agedeath

Figure 2.11 Portion of Data in File *agedeath*

	A
1	Age at Death (years)
2	83
3	76
4	78
5	74
6	35
7	78
8	73
9	84
10	55
11	73
12	35
13	78
14	65
15	81
16	109
17	91
18	87
19	76
20	80
21	78
22	85
23	81
24	73
25	42
26	53
27	54
28	98
29	50
30	85
31	77
32	81
33	89

Frequency Distributions for Quantitative Data

We can also create frequency distributions for quantitative data, but we must be more careful in defining the nonoverlapping bins to be used in the frequency distribution. For example, consider the case of an insurance actuary who is examining data related to the ages at death for 700 individuals. These data are contained the file *agedeath*, a portion of which is shown in Figure 2.11. The three steps necessary to define the classes for a frequency distribution with quantitative data are as follows:

1. Determine the number of nonoverlapping bins.
2. Determine the width of each bin.
3. Determine the range spanned by the set of bins.

We will next demonstrate these steps to create a frequency distribution for the age at death data.

Number of Bins Bins are formed by specifying the ranges used to group the data. As a general guideline, we recommend using from 5 to 20 bins. Using too many bins results in bins which contain only a few observations. With too many bins, the distribution does not capture generalizable patterns in the data. Using too few bins results in bins that aggregate data with too wide of range of values into the same bin. With too few bins, the distribution fails to accurately capture the variation in the data. For a small number of observations, as few as five or six bins may be used to summarize the data. For a larger number of observations, more bins are usually required. The determination of the number of bins is an inherently subjective decision, and the notion of a "best" number of bins depends on the subject matter and goal of the analysis. Because the number of observations in the *agedeath* file is relatively large ($n = 700$), we should choose a larger number of bins. We will use 16 bins to create the frequency distribution.

Width of the Bins As a general guideline, we recommend that the width be the same for each bin. Thus, the choices of the number of bins and the width of bins are not independent decisions. A larger number of bins means a smaller bin width and vice versa. To determine

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

an approximate bin width, we begin by identifying the largest and smallest data values. Then, with the desired number of bins specified, we can use the following expression to determine the approximate bin width.

Approximate Bin Width

$$\frac{\text{Largest data value} - \text{smallest data value}}{\text{Number of bins}} \tag{2.1}$$

The approximate bin width given by equation (2.1) can be rounded to a more convenient value based on the preference of the person developing the frequency distribution. For example, using equation (2.1) supplies an approximate bin width of $(109 - 0)/16 = 6.8125$. We round this number up to obtain a bin width of 7. By rounding up, we assure that 16 bins of width 7, with the first bin starting from the smallest data value, will cover the range of values in the data.

Range Spanned by Bins Once we have set the number of bins and the bin width, the remaining decision is to set the value at which the first bin begins. We must ensure that the set of bins spans the range of the data so that each observation belongs to exactly one bin. For example, a distribution with 16 bins and a bin width of 7 will cover a range of 112. Considering the data in the file *agedeath*, we observe that the smallest data value is 0 and the largest data value is 109. Because the range of data is 109, but the range of the bins is 112, we have four possible choices of the value to begin the first bin that would result in a set of bins that places each value from the data in a bin. We can define the first bin as [−3, 4] or [−2, 5] or [−1, 6] or [0, 7]. For example, if we begin the first bin at the smallest data value, the first bin is [0, 7], the second bin is (7, 14], the third bin is (14, 21]…, and the 16th bin is (105, 112]. In this case, the 16th bin's range extends past the largest data value of 109.

Using the data from the file *agedeath*, we select 0 years as the lower bin limit and 7 years as the upper bin limit for the first bin. This bin is denoted [0, 7] in Table 2.6. The brackets used in this bin label, "[" and "]", indicate that this bin is inclusive of the values 0 and 7. Sorting the age of death date in the file *agedeath* shows that there are 7 data values equal-to-or-larger than 0 and smaller-than-or-equal-to 7. These data values are 0, 0, 0, 0, 0, 1, 5. Therefore, the frequency associated with the first bin is 7. We then select 7 years as the lower bin limit and 14 years as the upper bin limit of the next bin. This bin is labeled as (7, 14]. The left parenthesis, "(", used here indicates that this bin does not include the value of 7, and the right bracket, "]", indicates that this bin does include the value of 14. Defining the bin in this manner ensures that this bin does not overlap with the first bin and an age at death value of 7 can only be included in a single bin, the bin [0, 7]. Returning to the age of death data and counting the number of data values that are greater than 7 but less-than-or-equal-to 14 shows that only a single data value, 10, falls into this bin. Therefore, the frequency associated with the bin (7, 14] is 1. We continue defining the lower and upper bin limits to obtain a total of 16 bins: [0, 7], (7, 14], (14, 21], (21, 28], (35, 42], (42, 49], (49, 56], (56, 63], (63, 70], (70, 77], (77, 84], (84, 91], (91, 98], (98, 105], (105, 112], (112+]. The final bin, "(112+] contains all values larger than 112 and is sometimes referred to as the "overflow" bin. Not all distributions contain an overflow bin, but they are often used for large data sets or data sets where there may be outliers.

Note that only the first bin is inclusive of both the lower and upper bin limits and the final bin includes all values larger than 112 years. All other bins are inclusive of the upper bin limit, but not the lower bin limit. This is a common way to define bins for a frequency distribution for quantitative values that ensures that each data value is only assigned to a single bin. The largest ages of death data value is 109 years which appears twice in the data set. These data values are assigned to the bin (105, 112]. With the number of bins, bin width, and range spanned by the bins determined, a frequency distribution can be obtained by counting the number of data values belonging to each bin. The complete frequency distribution, relative frequency distribution, and percent frequency distribution is shown in Table 2.6. We define the relative frequency and percent frequency distributions for quantitative data in the same manner as for qualitative data.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 2.6 Frequency, Relative Frequency, and Percent Frequency Distributions for the Age at Death Data

Age at Death Bin	Frequency	Relative Frequency	Percent Frequency (%)
[0, 7]	7	0.010	1.0
(7, 14]	1	0.001	0.1
(14, 21]	6	0.009	0.9
(21, 28]	7	0.010	1.0
(28, 35]	10	0.014	1.4
(35, 42]	15	0.021	2.1
(42, 49]	16	0.023	2.3
(49, 56]	36	0.051	5.1
(56, 63]	56	0.080	8.0
(63, 70]	78	0.111	11.1
(70, 77]	115	0.164	16.4
(77, 84]	148	0.211	21.1
(84, 91]	142	0.203	20.3
(91, 98]	58	0.083	8.3
(98, 105]	3	0.004	0.4
(105, 112]	2	0.003	0.3
(112+]	0	0.000	0.0

Using this frequency distribution, we can observe the following:

- The most common ages at death based on our bin definitions occurs for ages greater than 77 years and less-than-or-equal-to 84 years. 148 individuals, or 21.1% of the individuals in the sample, died at ages greater than 77 years and less-than-or-equal-to 84 years.
- Only 7 individuals, or 1% of the individuals in the sample, died at ages less-than-or-equal-to 7 years.
- No individuals in the sample lived for more than 112 years.

Other conclusions are possible, depending on the interests of the person viewing the frequency distribution. The value of a frequency distribution is that it provides insights about the data that are not easily obtained by viewing the data in their original unorganized form. Table 2.6 also shows the relative frequency distribution and percent frequency distribution for the age at death data.

In older versions of Excel, you must enter the FREQUENCY fuction as an array function by highlighting cells E2:E17 before entering the formula and then pressing ***Ctrl+Shift+Enter*** *rather than just* ***Enter****.*

Excel can also be used to create frequency distributions for quantitative data. In Figure 2.12, columns C and D contain the lower and upper limits defining the bins. We count the number of observations falling within each bin's range using the FREQUENCY function. In cell E2, we enter the formula $=FREQUENCY(A2{:}A701, D2{:}D17)$, where A2:A701 is the range for the data, and D2:D17 is the range containing the upper limits for each bin. After pressing **Enter** in cell E2, the range E2:E18 is populated with the number of observations falling within each bin's range.

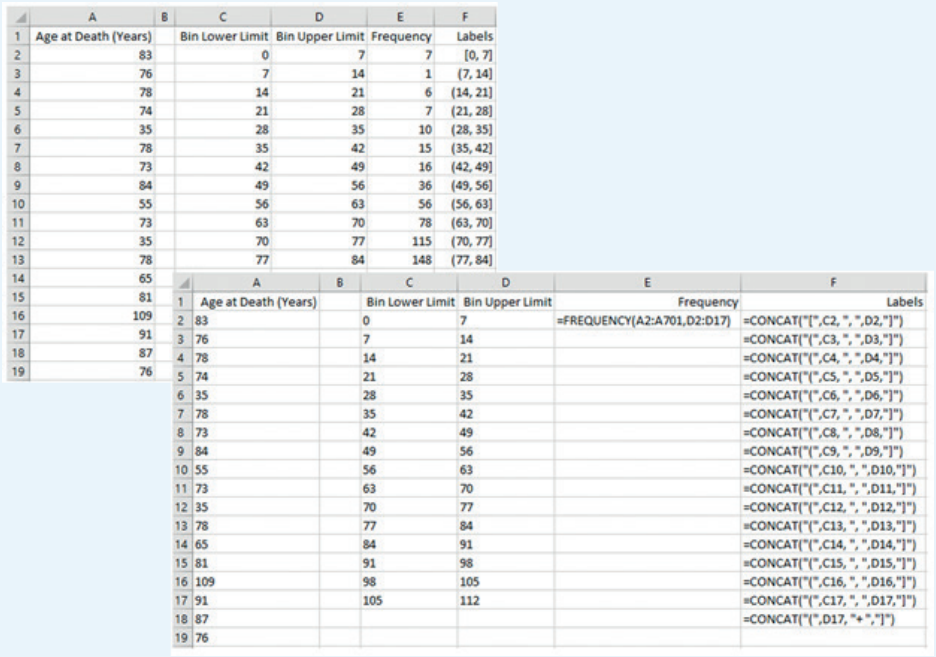
The frequencies shown in Figure 2.12 for the age at death data match those shown in Table 2.6

Histograms

A **histogram** is a graphical presentation of a frequency distribution, relative frequency distribution, or percent frequency distribution of quantitative data. Excel possesses functionality that defines each of the requirements for a frequency distribution (number of bins, bin

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.12 Frequency Distribution for the Age at Death Data Found using Excel



	A	B	C	D	E	F
1	Age at Death (Years)		Bin Lower Limit	Bin Upper Limit	Frequency	Labels
2	83		0	7	7	[0, 7]
3	76		7	14	1	(7, 14]
4	78		14	21	6	(14, 21]
5	74		21	28	7	(21, 28]
6	35		28	35	10	(28, 35]
7	78		35	42	15	(35, 42]
8	73		42	49	16	(42, 49]
9	84		49	56	36	(49, 56]
10	55		56	63	56	(56, 63]
11	73		63	70	78	(63, 70]
12	35		70	77	115	(70, 77]
13	78		77	84	148	(77, 84]
14	65					
15	81					
16	109					
17	91					
18	87					
19	76					

	A	B	C	D	E	F
1	Age at Death (Years)		Bin Lower Limit	Bin Upper Limit	Frequency	Labels
2	83		0	7	=FREQUENCY(A2:A701,D2:D17)	=CONCAT("[",C2, ", ",D2,"]")
3	76		7	14		=CONCAT("(",C3, ", ",D3,"]")
4	78		14	21		=CONCAT("(",C4, ", ",D4,"]")
5	74		21	28		=CONCAT("(",C5, ", ",D5,"]")
6	35		28	35		=CONCAT("(",C6, ", ",D6,"]")
7	78		35	42		=CONCAT("(",C7, ", ",D7,"]")
8	73		42	49		=CONCAT("(",C8, ", ",D8,"]")
9	84		49	56		=CONCAT("(",C9, ", ",D9,"]")
10	55		56	63		=CONCAT("(",C10, ", ",D10,"]")
11	73		63	70		=CONCAT("(",C11, ", ",D11,"]")
12	35		70	77		=CONCAT("(",C12, ", ",D12,"]")
13	78		77	84		=CONCAT("(",C13, ", ",D13,"]")
14	65		84	91		=CONCAT("(",C14, ", ",D14,"]")
15	81		91	98		=CONCAT("(",C15, ", ",D15,"]")
16	109		98	105		=CONCAT("(",C16, ", ",D16,"]")
17	91		105	112		=CONCAT("(",C17, ", ",D17,"]")
18	87					=CONCAT("(",D17, "+",")")
19	76					

limits, and range spanned by the bins) which it uses to automatically create a histogram. To demonstrate Excel's automatic histogram ability, consider again the data in the file *agedeath*. The following steps construct the histogram in Figure 2.13 illustrating the distribution of the ages at death.

Step 1. Select cells A1:A701
Step 2. Click the **Insert** tab on the Ribbon
Step 3. Click the **Insert Statistic Chart** button in the **Charts** group
When the list of statistic charts appears, select **Histogram**

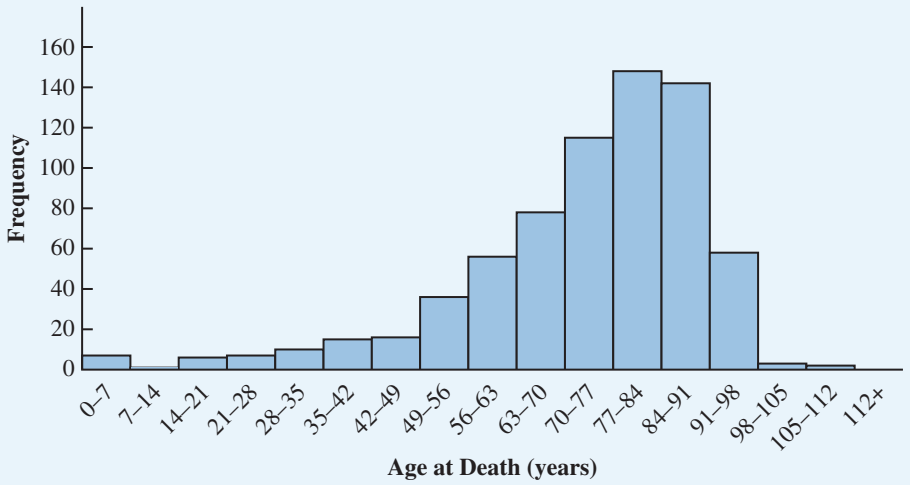
As Figure 2.13 illustrates, a histogram is simply a column chart with no spaces between the columns whose heights represent the frequencies of the corresponding bins. Eliminating the space between the columns allows a histogram to reflect the continuous nature of the variable of interest. For this data set, Excel automatically chose to use 16 bins, each spanning 7 years, traversing the range from 0 to 112.

Returning to Figure 2.13, we observe that the tallest column corresponds to the bin (77, 84]. To interpret this bin, we note that a square bracket indicates that the end value is included in the bin, and a round parenthesis indicates the end value is excluded. So the most common ages at death occur in the range greater than 77 years old and less than or equal to 84 years old. Furthermore, we observe the data are highly skewed left with most individuals dying at relatively old ages, but small numbers of individuals dying at young ages.

The choice of number of bins and bin width can strongly affect a histogram's display of a distribution. If an analyst prefers not to use the automatic histogram generated by Excel's Charts functionality, more user control is possible by using the Excel function FREQUENCY and a column chart to construct a histogram. Again, we will use the 700 observations in the file *agedeath* to manually create a histogram. The first step in manually creating a histogram is create a histogram by determining the number of bins, the bin limits, and the range spanned by the bins. We will use the same number of bins, bin limits, and

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.13 Histogram for the Age at Death Data Produced using Excel



range spanned by the bins for the age at death data as we did in generating the frequency distribution shown in Table 2.6. This approach using Excel is shown in Figure 2.12.

In older versions of Excel, the CONCAT function may not exist and you must use the CONCATENTATE function instead.

Using the data in D2:E17 of Figure 2.12, a histogram can then be used to illustrate the distribution of ages at death. In column F, we use the Excel CONCAT function to create a set of bin labels that we will use for the horizontal axis. The CONCAT function simply combines elements from different cells and/or different text stings into the same cell.

Step 1. Select cells D2:E17
Step 2. Click the **Insert** tab on the Ribbon
Step 3. Click the **Insert Column or Bar Chart** button in the **Charts** group
When the list of column and bar charts subtypes appears, click the **Clustered Column** button
Steps 1 through 3 result in a plotting both Bin Upper Limit and Frequency. To correct this, we need the following steps:
Step 4. Right click the chart and select **Change Chart Type…**
Step 5. When the **Change Chart Type** task pane appears, select the **Cluster Column** type that plots the appropriate number of variables (in this case, the single variable Frequency plotted with 16 columns)
Click **OK**
Step 6. Right click one of the columns in the chart and select **Format Data Series…**
When the **Format Data Series** task pane opens, click the **Series Options** button and set the **Gap Width** to 0%
Step 7. Right click the chart and choose **Select Data…**
When the **Select Data Source** dialog box appears, click **Edit** under **Horizontal (Category) Axis Labels**. In the **Axis Labels** dialog box, enter *=Data!F2$F$17* in the box below **Axis label range: and c**lick **OK**
Click **OK** to close the **Select Data Source** dialog box

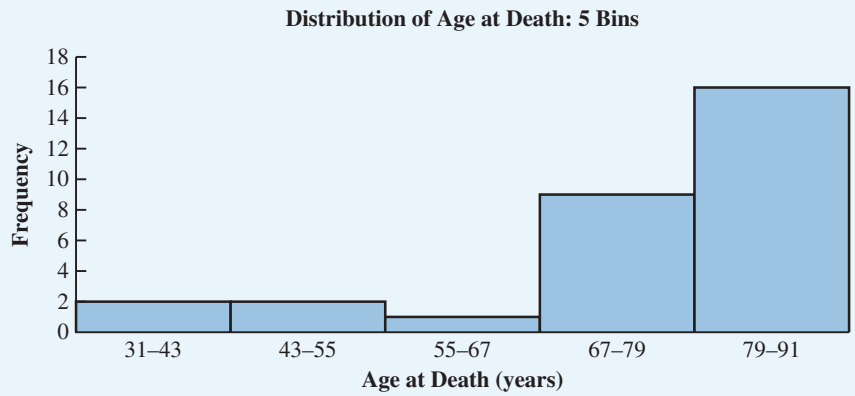
The preceding steps produce a histogram that matches the automatically generated histogram in Figure 2.13. However, using the Excel FREQUENCY command and the steps outlined here gives much greater control over the number of bins, the bin width, and the range spanned by the bins than automatically generating a histogram in Excel using the Insert Statistical Chart button.

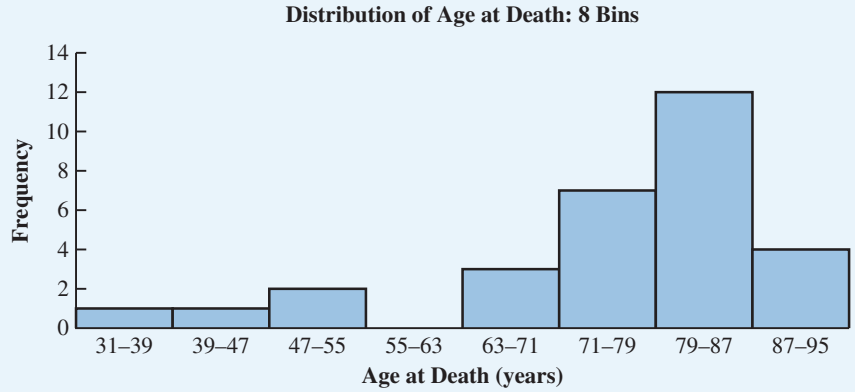
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

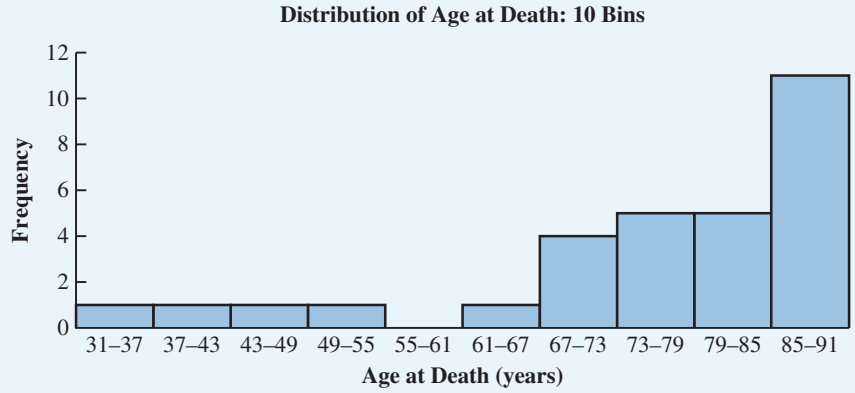
We note that the choice of the number of bins (and the corresponding bin width) may change the shape of the histogram (particularly for small data sets). Therefore, it is common to determine the number of bins and the appropriate bin width by trial and error. Once a possible number of bins are chosen, equation (2.1) is used to find the approximate bin width. The process can be repeated for several different numbers of bins.

To illustrate the impact of varying the number of bins and bin width on the shape of the histogram, we consider 30 observations in the *agedeath30* file, a subset of the data in the file *agedeath*. Figure 2.14 depicts three histograms, one using 5 bins with width 12, one

Figure 2.14 Histograms for the Age at Death Data using Different Numbers of Bins

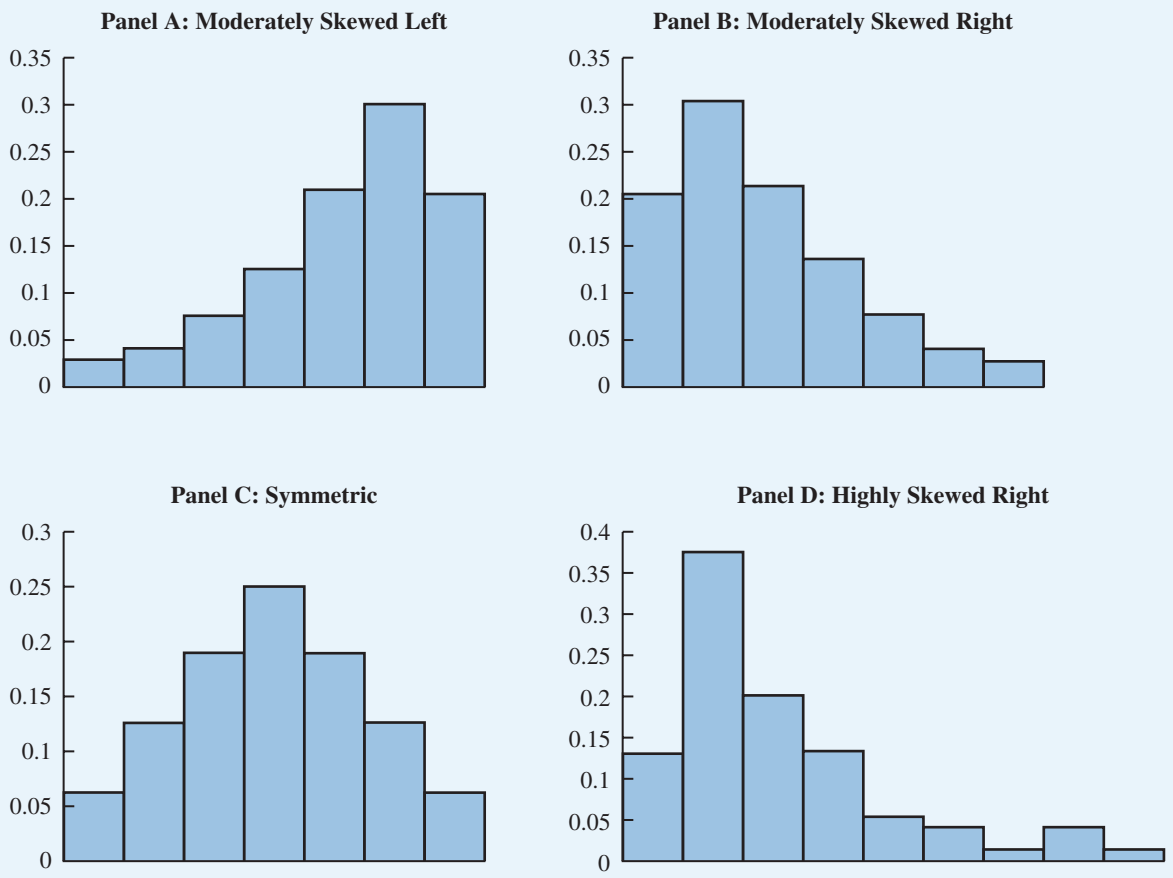






Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.15 Histograms Showing Distributions with Different Levels of Skewness



using 8 bins with width 8, and one using 10 bins with width 6. As Figure 2.14 illustrates, the choice of binning parameters can affect the shape of the histogram. The histogram with 5 bins and the histogram with 10 bins both suggest that the bin containing the oldest age range is the most likely. However, the histogram with 8 bins suggests that the bin containing the second oldest age range is most likely and the oldest age range is the third-most likely. An inspection of the data reveals that 6 out of the 30 observations had an age at death of 87 years, making the display of the distribution highly sensitive to the bin that contains 87.

One of the most important uses of a histogram is to provide information about the shape, or form, of a distribution. **Skewness**, or the lack of symmetry, is an important characteristic of the shape of a distribution. Figure 2.15 contains four histograms constructed from relative frequency distributions that exhibit different patterns of skewness. Figure 2.15 shows the histogram for a set of data moderately skewed to the left. A histogram is said to be skewed to the left if its tail extends farther to the left than to the right. This histogram is typical for exam scores, with no scores above 100%, most of the scores above 70%, and only a few really low scores.

As we have shown, histograms can be effective ways to visualize the distribution of a variable. However, when comparing the distribution of two or more variables, these columnar displays become cluttered. Next, we present a visualization tool that is particularly helpful for visualizing the distributions of multiple variables.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Frequency Polygons

A **frequency polygon** is useful for comparing distributions, particularly for quantitative variables. Like a histogram, a frequency polygon plots frequency counts of observations in a set of bins. However, a frequency polygon uses lines to connect the counts of different bins, in contrast to a histogram, which uses columns to depict the counts in different bins. To demonstrate the construction of histograms and frequency polygons for two different variables, we consider the data in the file *agedeath_sex*, which supplements the age at death information for the 700 individuals in the file *agedeath* with the self-identified sex of each of these individuals. Similar to how we constructed the frequency distribution for all 700 observations, we must create separate frequency distributions for the female and male observations, respectively. However, when comparing frequency distributions, it is a good practice to use relative frequency calculations because the total number of observations in the two distributions may not be the same. For instance, in the file *agedeath_sex*, there are 327 self-identified female observations and 373 self-identified male observations, so comparing just the count of observations in the bins may distort the comparison.

DATA*file*
agedeath_sex

In Figure 2.16, we compute the number of self-identified female and male observations by entering the formula *=COUNTIF(B2:B701, "Female")* in cell F3 and the formula *=COUNTIF(B2:B701, "Male")* in cell G3. We use these counts to compute the relative

Figure 2.16 Creating Frequency Distributions for Female and Male Observations in the File *agedeath_sex*

	A	B	C	D	E	F	G	H
1	Age at Death (Years)	Sex						
2	76	Female				Female	Male	
3	35	Female			Sample Size	=COUNTIF(B2:B701,"Female")	=COUNTIF(B2:B701,"Male")	
4	84	Female						
5	55	Female		Bin Lower Limit	Bin Upper Limit	Female	Male	Labels
6	35	Female		0	7	=FREQUENCY(A2:A328,E6:E21)/$F$3	=FREQUENCY(A329:A701,E6:E21)/$G$3	=CONCAT("[",D6,",",E6,"]")
7	81	Female		7	14			=CONCAT("(",D7,",",E7,"]")
8	109	Female		14	21			=CONCAT("(",D8,",",E8,"]")
9	91	Female		21	28			=CONCAT("(",D9,",",E9,"]")
10	87	Female		28	35			=CONCAT("(",D10,",",E10,"]")
11	42	Female		35	42			=CONCAT("(",D11,",",E11,"]")
12	50	Female		42	49			=CONCAT("(",D12,",",E12,"]")
13	85	Female		49	56			=CONCAT("(",D13,",",E13,"]")
14	89	Female		56	63			=CONCAT("(",D14,",",E14,"]")
15	67	Female		63	70			=CONCAT("(",D15,",",E15,"]")
16	84	Female		70	77			=CONCAT("(",D16,",",E16,"]")
17	60	Female						=CONCAT("(",D17,",",E17,"]")
18	65	Female						=CONCAT("(",D18,",",E18,"]")
19	86	Female						=CONCAT("(",D19,",",E19,"]")
20	60	Female						=CONCAT("(",D20,",",E20,"]")
21	77	Female						=CONCAT("(",D21,",",E21,"]")
22	74	Female						=CONCAT("(",E21,"+","]")
23	73	Female						

	A	B	C	D	E	F	G	H
1	Age at Death (Years)	Sex						
2	76	Female				Female	Male	
3	35	Female			Sample Size	327	373	
4	84	Female						
5	55	Female		Bin Lower Limit	Bin Upper Limit	Female	Male	Labels
6	35	Female		0	7	0.009	0.011	[0,7]
7	81	Female		7	14	0.000	0.003	(7,14]
8	109	Female		14	21	0.009	0.008	(14,21]
9	91	Female		21	28	0.003	0.016	(21,28]
10	87	Female		28	35	0.015	0.013	(28,35]
11	42	Female		35	42	0.012	0.029	(35,42]
12	50	Female		42	49	0.012	0.032	(42,49]
13	85	Female		49	56	0.040	0.062	(49,56]
14	89	Female		56	63	0.083	0.078	(56,63]
15	67	Female		63	70	0.095	0.126	(63,70]
16	84	Female		70	77	0.159	0.169	(70,77]
17	60	Female		77	84	0.168	0.249	(77,84]
18	65	Female		84	91	0.251	0.161	(84,91]
19	86	Female		91	98	0.128	0.043	(91,98]
20	60	Female		98	105	0.009	0.000	(98,105]
21	77	Female		105	112	0.006	0.000	(105,112]
22	74	Female				0.000	0.000	(112+]
23	73	Female						

DATA*file*
agedeath_sexfrequency

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

frequencies of male and female observations in the bins, which are defined by the lower and upper limits in Columns D and E. For the female and male observations (which have already been sorted), we count the number of observations falling within each bin's range using the FREQUENCY function. In cell F6, we enter the formula *=FREQUENCY (A2:A328,E6:E21)/$F$3*, where A2:A328 is the range for the female observations, E6:E21 is the range containing the upper limits for each bin, and cell F3 contains the total number of female observations. After hitting enter in cell F6, the range F6:F22 is populated with the relative frequency of female observations falling within each bin's range. In cell G6, we enter the formula *=FREQUENCY(A329:A701,E6:E21)/$G$3*, where A329:A701 is the range for the male observations, E6:E21 is the range containing the upper limits for each bin, and cell G3 contains the total number of male observations. After hitting Enter in cell G6, the range G6:G21 is populated with the relative frequency of male observations falling within each bin's range.

The following steps produce a chart with histograms for the male and female observations using a clustered column chart as shown in Figure 2.17a.

Clustered column charts are covered more extensively in Chapter 3.

Step 1. Select cells F6:G21
Step 2. Click the **Insert** tab on the Ribbon
Step 3. Click the **Insert Column or Bar Chart** button in the **Charts** group
Step 4. When the list of column and bar chart subtypes appears, click the **Clustered Column** button

To display the age at death distributions for female and male observations as a frequency polygon as shown in Figure 2.17b, we execute the following steps.

Step 1. Select cells F6:G21
Step 2. Click the **Insert** tab on the Ribbon
Step 3. Click the **Insert Line or Area Chart** button in the **Charts** group
When the list of line and area chart subtypes appears, click the **Line** button

As Figure 2.17 illustrates for the age at death distributions for males and females, the use of lines in the frequency polygons preserves the continuity of each distribution's shape, while the use of a clustered column chart provides a disrupted visualization of the distributions. While frequency polygons provide for a more transparent comparison of two or more distributions, for a single distribution they do not support the magnitude comparison of different bins as well as a histogram. Therefore, histograms are typically preferred for the visualization of a single variable's distribution.

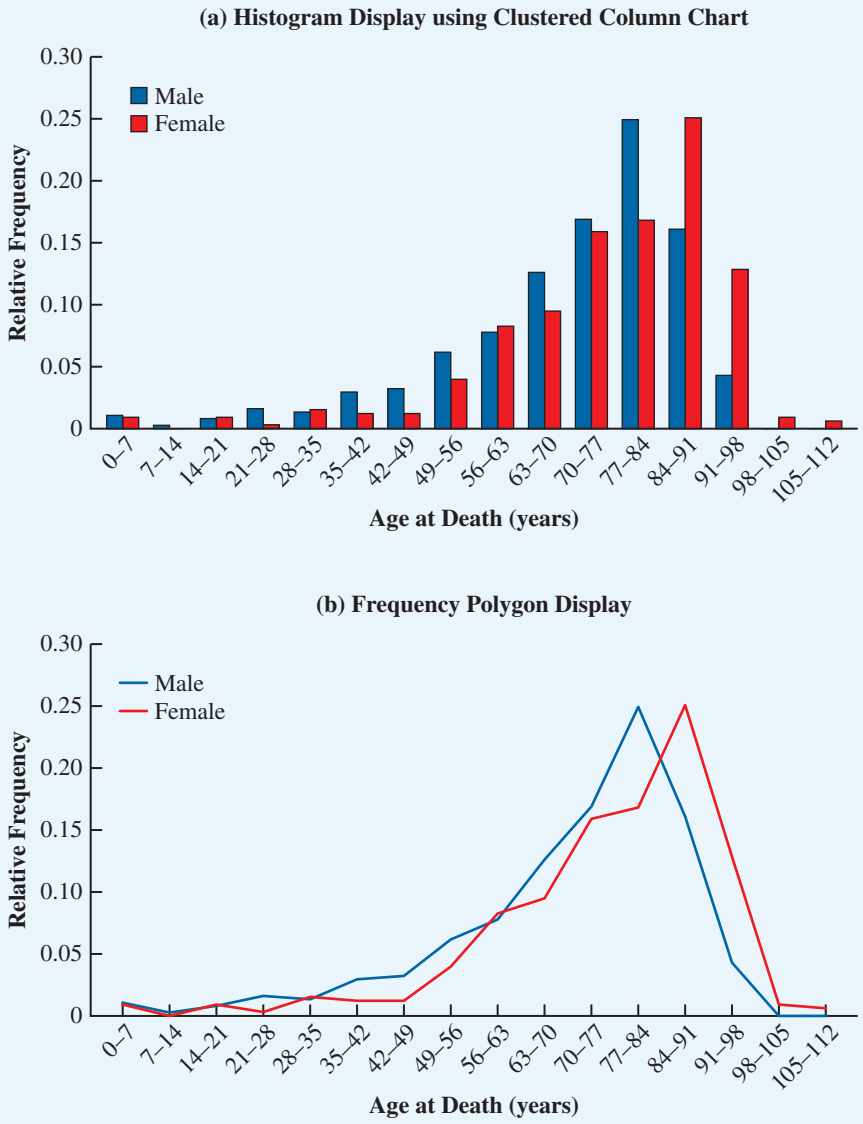
Cumulative Distributions

A variation of the frequency distribution that provides another tabular summary of quantitative data is the **cumulative frequency distribution**, which uses the number of bins, bin widths, and range spanned by the bins developed for the frequency distribution. However, rather than showing the frequency of each bin, the cumulative frequency distribution shows the number of data items with values less than or equal to the upper limit of each bin. The first two columns of Table 2.7 provide the cumulative frequency distribution for the age at death data.

To understand how the cumulative frequencies are determined, consider the row of Table 2.7 with the description "Less than or equal to 21." The cumulative frequency for this row is simply the sum of the frequencies for all bins with data values less than or equal to 21. For the frequency distribution in Table 2.6, the sum of the frequencies for bins (0, 7], (7, 14], and (14, 21] is $7 + 1 + 6 = 14$ data values are less than or equal to 21. Hence, the cumulative frequency for this row is 14. Therefore, the cumulative frequency distribution in Table 2.7 shows that 14 individuals died at age 21 years or less. Table 2.7 also shows that 31 individuals died at 35 years of age or less, 46 individuals died at age 42 year or less, and so on.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.17 Comparing Frequency Polygon to Clustered Column Display of Distribution of Age at Death Data by Sex



As a final point, a cumulative relative frequency distribution shows the proportion of data items, and a cumulative percent frequency distribution shows the percentage of data items with values less than or equal to the upper limit of each bin. The cumulative relative frequency distribution can be computed either by summing the relative frequencies in the relative frequency distribution or by dividing the cumulative frequencies by the total number of items. For instance, we find the cumulative relative frequencies in column 3 of Table 2.7 by dividing the cumulative frequencies in column 2 by the total number of items ($n = 700$). The cumulative percent frequencies were again computed by multiplying the relative frequencies by 100. The cumulative relative and percent frequency distributions show that 0.707 of the individuals, or 70.7%, died at age 84 or less, and 0.993 of the individuals, or 99.3%, died at age 98 or less.

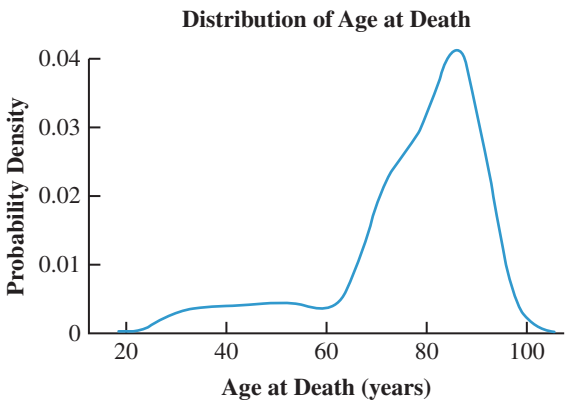
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 2.7 Cumulative Frequency, Cumulative Relative Frequency, and Cumulative Percent Frequency Distributions for the Age at Death Data

Age at Death (years)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency (%)
Less than or equal to 7	7	0.010	1.0
Less than or equal to 14	8	0.011	1.1
Less than or equal to 21	14	0.020	2.0
Less than or equal to 28	21	0.030	3.0
Less than or equal to 35	31	0.044	4.4
Less than or equal to 42	46	0.066	6.6
Less than or equal to 49	62	0.089	8.9
Less than or equal to 56	98	0.140	14.0
Less than or equal to 63	154	0.220	22.0
Less than or equal to 70	232	0.331	33.1
Less than or equal to 77	347	0.496	49.6
Less than or equal to 84	495	0.707	70.7
Less than or equal to 91	637	0.910	91.0
Less than or equal to 98	695	0.993	99.3
Less than or equal to 105	698	0.997	99.7
Less than or equal to 112	700	1.000	100.0

Notes + Comments

1. Due to rounding, summing up the relative frequencies shown in Table 2.6 may generate slightly different values than those shown in Table 2.7 for the cumulative relative frequencies. In general, it is recommended that values not be rounded when calculating the cumulative relative frequencies which is best accomplished using Excel or another software package.
2. Distributions are often used when discussing concepts related to probability and simulation because they are used to describe uncertainty. In Chapter 5, we will discuss probability distributions, and then in Chapter 13 we will revisit distributions when we introduce simulation models.
3. A kernel density chart is a "continuous" alternative to a histogram designed to overcome the reliance of histograms on the choice of number of bins and bin width. Kernel density charts employ a smoothing technique known as kernel density estimation to generate a more robust visualization of the distribution of a set of values. As an example, a kernel density chart for the 30 observations in *agedeath30* file is displayed with this note. Comparing the kernel density chart to the histograms in Figure 2.14, we observe that the kernel density chart smooths the extremes of the histograms in an attempt to generalize the patterns in the data. Excel does not have built-in functionality to construct a kernel density chart (which is not the same as a frequency polygon), but many statistical software packages such as R do.



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2.5 Measures of Location

Mean (Arithmetic Mean)

The most commonly used measure of location is the **mean (arithmetic mean)**, or average value, for a variable. The mean provides a measure of central location for the data. If the data are for a sample (typically the case), the mean is denoted by $\bar{x}$. The sample mean is a point estimate of the (typically unknown) population mean for the variable of interest. If the data for the entire population are available, the population mean is computed in the same manner, but denoted by the Greek letter μ.

In statistical formulas, it is customary to denote the value of variable x for the first observation by x_1, the value of variable x for the second observation by x_2, and so on. In general, the value of variable x for the ith observation is denoted by x_i. For a sample with n observations, the formula for the sample mean is as follows.

If the data set is not a sample, but is the entire population with N observations, the population mean is computed directly by: $\mu = \frac{\Sigma x_i}{N}$.

Sample Mean

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{2.2}$$

To illustrate the computation of a sample mean, suppose a sample of home sales is taken for a suburb of Cincinnati, Ohio. Table 2.8 shows the collected data. The mean home selling price for the sample of 12 home sales is

$$\begin{aligned} \bar{x} &= \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12} \\ &= \frac{138{,}000 + 254{,}000 + \cdots + 456{,}250}{12} \\ &= \frac{2{,}639{,}250}{12} = 219{,}937.50 \end{aligned}$$

The mean can be found in Excel using the AVERAGE function. Figure 2.18 shows the Home Sales data from Table 2.8 in an Excel spreadsheet. The value for the mean in cell E2 is calculated using the formula $=AVERAGE(B2:B13)$.

DATA*file*
homesales

Table 2.8 Data on Home Sales in a Cincinnati, Ohio, Suburb

Home Sale	Selling Price ($)
1	138,000
2	254,000
3	186,000
4	257,500
5	108,000
6	254,000
7	138,000
8	298,000
9	199,500
10	208,000
11	142,000
12	456,250

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.18 Calculating the Mean, Median, and Modes for the Home Sales Data Using Excel

	A	B	C	D	E
1	**Home Sale**	**Selling Price ($)**			
2	1	138000		**Mean:**	=AVERAGE(B2:B13)
3	2	254000		**Median:**	=MEDIAN(B2:B13)
4	3	186000		**Mode 1:**	=MODE.MULT(B2:B13)
5	4	257500		**Mode 2:**	
6	5	108000			
7	6	254000			
8	7	138000			
9	8	298000			
10	9	199500			
11	10	208000			
12	11	142000			
13	12	456250			

	A	B	C	D	E
1	**Home Sale**	**Selling Price ($)**			
2	1	138,000		**Mean:**	$219,937.50
3	2	254,000		**Median:**	$203,750.00
4	3	186,000		**Mode 1:**	$138,000.00
5	4	257,500		**Mode 2:**	$254,000.00
6	5	108,000			
7	6	254,000			
8	7	138,000			
9	8	298,000			
10	9	199,500			
11	10	208,000			
12	11	142,000			
13	12	456,250			

Median

The **median**, another measure of central location, is the value in the middle when the data are arranged in ascending order (smallest to largest value). With an odd number of observations, the median is the middle value. An even number of observations has no single middle value. In this case, we follow convention and define the median as the average of the values for the middle two observations.

Let us apply this definition to compute the median class size for a sample of five college classes. Arranging the data in ascending order provides the following list:

$$32 \quad 42 \quad 46 \quad 46 \quad 54$$

Because $n = 5$ is odd, the median is the middle value. Thus, the median class size is 46 students. Even though this data set contains two observations with values of 46, each observation is treated separately when we arrange the data in ascending order.

Suppose we also compute the median value for the 12 home sales in Table 2.8. We first arrange the data in ascending order.

$$108{,}000 \quad 138{,}000 \quad 138{,}000 \quad 142{,}000 \quad 186{,}000 \quad \underbrace{199{,}500 \quad 208{,}000}_{\text{Middle Two Values}} \quad 254{,}000 \quad 254{,}000 \quad 257{,}500 \quad 298{,}000 \quad 456{,}250$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Because $n = 12$ is even, the median is the average of the middle two values: 199,500 and 208,000.

$$\text{Median} = \frac{199{,}500 + 208{,}000}{2} = 203{,}750$$

The median of a data set can be found in Excel using the function MEDIAN. In Figure 2.18, the value for the median in cell E3 is found using the formula *=MEDIAN(B2:B13)*.

Although the mean is the more commonly used measure of central location, in some situations the median is preferred. The mean is influenced by extremely small and large data values. Notice that the median is smaller than the mean in Figure 2.18. This is because the one large value of $456,250 in our data set inflates the mean but does not have the same effect on the median. Notice also that the median would remain unchanged if we replaced the $456,250 with a sales price of $1.5 million. In this case, the median selling price would remain $203,750, but the mean would increase to $306,916.67. If you were looking to buy a home in this suburb, the median gives a better indication of the central selling price of the homes there. We can generalize, saying that whenever a data set contains extreme values or is severely skewed, the median is often the preferred measure of central location.

Mode

A third measure of location, the **mode**, is the value that occurs most frequently in a data set. To illustrate the identification of the mode, consider the sample of five class sizes.

32 42 46 46 54

The only value that occurs more than once is 46. Because this value, occurring with a frequency of 2, has the greatest frequency, it is the mode. To find the mode for a data set with only one most often occurring value in Excel, we use the MODE.SNGL function.

Occasionally the greatest frequency occurs at two or more different values, in which case more than one mode exists. If data contain at least two modes, we say that they are *multimodal*. A special case of multimodal data occurs when the data contain exactly two modes; in such cases we say that the data are *bimodal*. In multimodal cases when there are more than two modes, the mode is almost never reported because listing three or more modes is not particularly helpful in describing a location for the data. Also, if no value in the data occurs more than once, we say the data have no mode.

The Excel MODE.SNGL function will return only a single most-often-occurring value. For multimodal distributions, we must use the MODE.MULT command in Excel to return more than one mode. For example, two selling prices occur twice in Table 2.8: $138,000 and $254,000. Hence, these data are bimodal. To find both of the modes in Excel, we take these steps:

In older versions of Excel, you may have to first select cells E4 and E5 in step 1 and then press CTRL+SHIFT+ENTER rather than just ENTER in step 2 to display both modes.

Step 1. Select cell E4
Step 2. Type the formula *=MODE.MULT(B2:B13)* and press **Enter**

Excel enters the values for both modes of this data set in cells E4 and E5: $138,000 and $254,000.

Geometric Mean

The **geometric mean** is a measure of location that is calculated by finding the nth root of the product of n values. The general formula for the sample geometric mean, denoted $\bar{x}_g$, follows.

The geometric mean for a population is computed similarly but is defined as μ_g to denote that it is computed using the entire population.

SAMPLE GEOMETRIC MEAN

$$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\cdots(x_n)} = [(x_1)(x_2)\cdots(x_n)]^{1/n} \qquad \textbf{(2.3)}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The geometric mean is often used in analyzing growth rates in financial data. In these types of situations, the arithmetic mean or average value will provide misleading results.

To illustrate the use of the geometric mean, consider Table 2.9, which shows the percentage annual returns, or growth rates, for a mutual fund over the past 10 years. Suppose we want to compute how much \$100 invested in the fund at the beginning of year 1 would be worth at the end of year 10. We start by computing the balance in the fund at the end of year 1. Because the percentage annual return for year 1 was −22.1%, the balance in the fund at the end of year 1 would be

$$\$100 - 0.221(\$100) = \$100(1 - 0.221) = \$100(0.779) = \$77.90$$

The growth factor for each year is 1 plus 0.01 times the percentage return. A growth factor less than 1 indicates negative growth, whereas a growth factor greater than 1 indicates positive growth. The growth factor cannot be less than zero.

We refer to 0.779 as the **growth factor** for year 1 in Table 2.9. We can compute the balance at the end of year 1 by multiplying the value invested in the fund at the beginning of year 1 by the growth factor for year 1: \$100(0.779) = \$77.90.

The balance in the fund at the end of year 1, \$77.90, now becomes the beginning balance in year 2. So, with a percentage annual return for year 2 of 28.7%, the balance at the end of year 2 would be

$$\$77.90 + 0.287(\$77.90) = \$77.90(1 + 0.287) = \$77.90(1.287) = \$100.26$$

Note that 1.287 is the growth factor for year 2. By substituting \$100(0.779) for \$77.90, we see that the balance in the fund at the end of year 2 is

$$\$100(0.779)(1.287) = \$100.26$$

In other words, the balance at the end of year 2 is just the initial investment at the beginning of year 1 times the product of the first two growth factors. This result can be generalized to show that the balance at the end of year 10 is the initial investment times the product of all 10 growth factors.

$$\$100[(0.779)(1.287)(1.109)(1.049)(1.158)(1.055)(0.630)(1.265)(1.151)(1.021)]$$
$$= \$100(1.335) = \$133.45$$

So a \$100 investment in the fund at the beginning of year 1 would be worth \$133.45 at the end of year 10. Note that the product of the 10 growth factors is 1.335. Thus, we can compute the balance at the end of year 10 for any amount of money invested at the beginning of year 1 by multiplying the value of the initial investment by 1.335. For instance, an initial investment of \$2,500 at the beginning of year 1 would be worth \$2,500(1.335), or approximately \$3,337.50, at the end of year 10.

Table 2.9 Percentage Annual Returns and Growth Factors for the Mutual Fund Data

Year	Return (%)	Growth Factor
1	−22.1	0.779
2	28.7	1.287
3	10.9	1.109
4	4.9	1.049
5	15.8	1.158
6	5.5	1.055
7	−37.0	0.630
8	26.5	1.265
9	15.1	1.151
10	2.1	1.021

DATA*file*
mutualfundreturns

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

What was the mean percentage annual return or mean rate of growth for this investment over the 10-year period? The geometric mean of the 10 growth factors can be used to answer this question. Because the product of the 10 growth factors is 1.335, the geometric mean is the 10th root of 1.335, or

$$\bar{x}_g = \sqrt[10]{1.335} = 1.029$$

The geometric mean tells us that annual returns grew at an average annual rate of (1.029 − 1) 100, or 2.9%. In other words, with an average annual growth rate of 2.9%, a $100 investment in the fund at the beginning of year 1 would grow to $\$100(1.029)^{10} = \$133.09$ at the end of 10 years. We can use Excel to calculate the geometric mean for the data in Table 2.9 by using the function GEOMEAN. In Figure 2.19, the value for the geometric mean in cell C13 is found using the formula *=GEOMEAN(C2:C11)*.

It is important to understand that the arithmetic mean of the percentage annual returns does not provide the mean annual growth rate for this investment. The sum of the 10 percentage annual returns in Table 2.9 is 50.4. Thus, the arithmetic mean of the 10 percentage returns is 50.4/10 = 5.04%. A salesperson might try to convince you to invest in this fund by stating that the mean annual percentage return was 5.04%. Such a statement is not only misleading, it is inaccurate. A mean annual percentage return of 5.04% corresponds to an average growth factor of 1.0504. So, if the average growth factor were really 1.0504, $100 invested in the fund at the beginning of year 1 would have grown to $\$100(1.0504)^{10} = \$163.51$ at the end of 10 years. But, using the 10 annual percentage returns in Table 2.9, we showed that an initial $100 investment is worth $133.09 at the end of 10 years. The salesperson's claim that the mean annual percentage return is 5.04% grossly overstates the true growth for this mutual fund. The problem is that the arithmetic mean is appropriate only for an additive process. For a multiplicative process, such as applications involving growth rates, the geometric mean is the appropriate measure of location.

While the application of the geometric mean to problems in finance, investments, and banking is particularly common, the geometric mean should be applied any time you want to determine the mean rate of change over several successive periods. Other common applications include changes in the populations of species, crop yields, pollution levels, and birth and death rates. The geometric mean can also be applied to changes that occur over any

Figure 2.19 Calculating the Geometric Mean for the Mutual Fund Data Using Excel

	A	B	C
1	Year	Return (%)	Growth Factor
2	1	-22.1	0.779
3	2	28.7	1.287
4	3	10.9	1.109
5	4	4.9	1.049
6	5	15.8	1.158
7	6	5.5	1.055
8	7	-37	0.63
9	8	26.5	1.265
10	9	15.1	1.151
11	10	2.1	1.021
12			
13		Geometric Mean:	=GEOMEAN(C2:C11)

	A	B	C
1	Year	Return (%)	Growth Factor
2	1	-22.1	0.779
3	2	28.7	1.287
4	3	10.9	1.109
5	4	4.9	1.049
6	5	15.8	1.158
7	6	5.5	1.055
8	7	-37	0.63
9	8	26.5	1.265
10	9	15.1	1.151
11	10	2.1	1.021
12			
13		Geometric Mean:	1.029

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

number of successive periods of any length. In addition to annual changes, the geometric mean is often applied to find the mean rate of change over quarters, months, weeks, and even days.

2.6 Measures of Variability

In addition to measures of location, it is often desirable to consider measures of variability, or dispersion. For example, suppose that you are considering two financial funds. Both funds require a $1,000 annual investment. Table 2.10 shows the annual payouts for Fund A and Fund B for $1,000 investments over the past 20 years. Fund A has paid out exactly $1,100 each year for an initial $1,000 investment. Fund B has had many different payouts, but the mean payout over the previous 20 years is also $1,100. But would you consider the payouts of Fund A and Fund B to be equivalent? Clearly, the answer is no. The difference between the two funds is due to variability.

Figure 2.20 shows a histogram for the payouts received from Funds A and B. Although the mean payout is the same for the two funds, their histograms differ in that the payouts associated with Fund B have greater variability. Sometimes the payouts are considerably larger than the mean, and sometimes they are considerably smaller. In this section, we present several different ways to measure variability.

Range

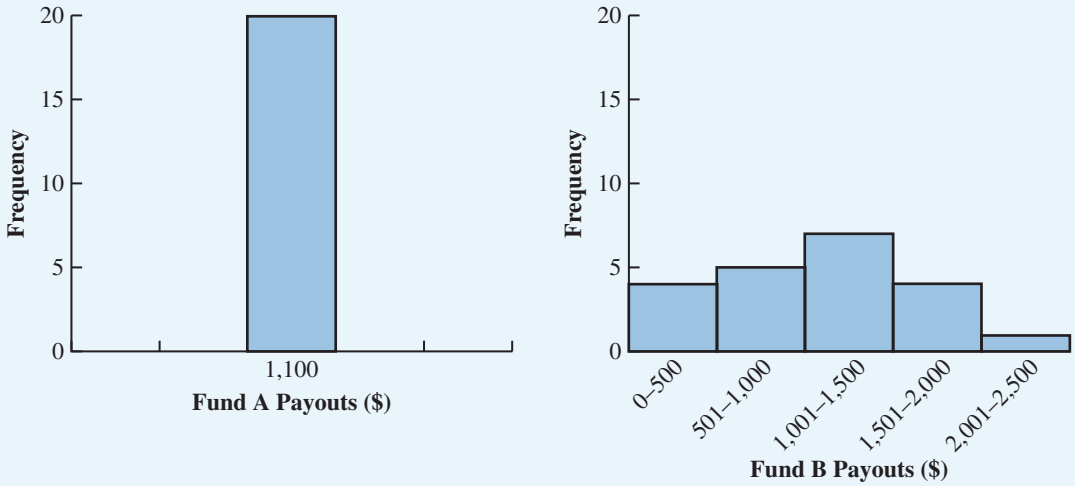
The simplest measure of variability is the **range**. The range can be found by subtracting the smallest value from the largest value in a data set. Let us return to the home sales data set to demonstrate the calculation of range. Refer to the data from home sales prices in Table 2.8. The largest home sales price is $456,250, and the smallest is $108,000. The range is $456,250 − $108,000 = $348,250.

Table 2.10 Annual Payouts for Two Different Investment Funds

Year	Fund A ($)	Fund B ($)
1	1,100	700
2	1,100	2,500
3	1,100	1,200
4	1,100	1,550
5	1,100	1,300
6	1,100	800
7	1,100	300
8	1,100	1,600
9	1,100	1,500
10	1,100	350
11	1,100	460
12	1,100	890
13	1,100	1,050
14	1,100	800
15	1,100	1,150
16	1,100	1,200
17	1,100	1,800
18	1,100	100
19	1,100	1,750
20	1,100	1,000
Mean	1,100	1,100

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.20 Histograms for Payouts of Past 20 Years from Fund A and Fund B



Although the range is the easiest of the measures of variability to compute, it is seldom used as the only measure. The reason is that the range is based on only two of the observations and thus is highly influenced by extreme values. If, for example, we replace the selling price of \$456,250 with \$1.5 million, the range would be \$1,500,000 − \$108,000 = \$1,392,000. This large value for the range would not be especially descriptive of the variability in the data because 11 of the 12 home selling prices are between \$108,000 and \$298,000.

The range can be calculated in Excel using the MAX and MIN functions. The range value in cell E7 of Figure 2.21 calculates the range using the formula *=MAX(B2:B13) – MIN(B2:B13)*. This subtracts the smallest value in the range B2:B13 from the largest value in the range B2:B13.

Variance

The **variance** is a measure of variability that utilizes all the data. The variance is based on the *deviation about the mean*, which is the difference between the value of each observation (x_i) and the mean. For a sample, a deviation of an observation about the mean is written ($x_i - \bar{x}$). In the computation of the variance, the deviations about the mean are squared.

If the data are for a population, the population variance, σ^2, can be computed directly (rather than estimated by the sample variance). For a population of N *observations and with* μ *denoting the population mean, population variance is computed by* $\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$.

In most statistical applications, the data being analyzed are for a sample. When we compute a sample variance, we are often interested in using it to estimate the population variance, σ^2. Although a detailed explanation is beyond the scope of this text, for a random sample, it can be shown that, if the sum of the squared deviations about the sample mean is divided by $n - 1$, and not n, the resulting sample variance provides an unbiased estimate of the population variance.[1]

For this reason, the sample variance, denoted by s^2, is defined as follows:

Sample Variance

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \tag{2.4}$$

[1]Unbiased means that if we take a large number of independent random samples of the same size from the population and calculate the sample variance for each sample, the average of these sample variances will tend to be equal to the population variance.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.21 Adding Variability Measures to the Mean, Median, and Modes Measures for the Home Sales Data in Excel

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138000		Mean:	=AVERAGE(B2:B13)
3	2	254000		Median:	=MEDIAN(B2:B13)
4	3	186000		Mode 1:	=MODE.MULT(B2:B13)
5	4	257500		Mode 2:	
6	5	108000			
7	6	254000		Range:	=MAX(B2:B13)-MIN(B2:B13)
8	7	138000		Variance:	=VAR.S(B2:B13)
9	8	298000		Standard Deviation:	=STDEV.S(B2:B13)
10	9	199500			
11	10	208000		Coefficient of Variation:	=E9/E2
12	11	142000			
13	12	456250		85th Percentile:	=PERCENTILE.EXC(B2:B13,0.85)
14					
15				1st Quartile:	=QUARTILE.EXC(B2:B13,1)
16				2nd Quartile:	=QUARTILE.EXC(B2:B13,2)
17				3rd Quartile:	=QUARTILE.EXC(B2:B13,3)
18					
19				IQR:	=E17-E15

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138,000		Mean:	$ 219,937.50
3	2	254,000		Median:	$ 203,750.00
4	3	186,000		Mode 1:	$ 138,000.00
5	4	257,500		Mode 2:	$ 254,000.00
6	5	108,000			
7	6	254,000		Range:	$ 348,250.00
8	7	138,000		Variance:	9037501420
9	8	298,000		Standard Deviation:	$ 95,065.77
10	9	199,500			
11	10	208,000		Coefficient of Variation:	43.22%
12	11	142,000			
13	12	456,250		85th Percentile:	$ 305,912.50
14					
15				1st Quartile:	$ 139,000.00
16				2nd Quartile:	$ 203,750.00
17				3rd Quartile:	$ 256,625.00
18					
19				IQR:	$ 117,625.00

To illustrate the computation of the sample variance, we will use the data on class size from page 41 for the sample of five college classes. A summary of the data, including the computation of the deviations about the mean and the squared deviations about the mean, is shown in Table 2.11. The sum of squared deviations about the mean is $\Sigma(x_i - \bar{x})^2 = 256$. Hence, with $n - 1 = 4$, the sample variance is

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$

Note that the units of variance are squared. For instance, the sample variance for our calculation is $s^2 = 64\,(\text{students})^2$. In Excel, you can find the variance for sample data using the VAR.S function. Figure 2.21 shows the data for home sales examined in the previous section. The variance in cell E8 is calculated using the formula $=VAR.S(B2:B13)$. Excel calculates the variance of the sample of 12 home sales to be 9,037,501,420.

Standard Deviation

The **standard deviation** is defined to be the positive square root of the variance. We use s to denote the sample standard deviation and σ to denote the population standard deviation. The sample standard deviation, s, is a point estimate of the population standard deviation, σ, and is derived from the sample variance in the following way:

Sample Standard Deviation

$$s = \sqrt{s^2} \tag{2.5}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 2.11 Computation of Deviations and Squared Deviations About the Mean for the Class Size Data

Number of Students in Class (x_i)	Mean Class Size ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		$\Sigma(x_i - \bar{x}) = 0$	$\Sigma(x_i - \bar{x})^2 = 256$

If the data are for a population, the population standard deviation σ is obtained by taking the positive square root of the population variance: $\sigma = \sqrt{\sigma^2}$. To calculate the population variance and population stan dard deviation in Excel, we use the functions =VAR.P *and* =STDEV.P.

The sample variance for the sample of class sizes in five college classes is $s^2 = 64$. Thus, the sample standard deviation is $s = \sqrt{64} = 8$.

Recall that the units associated with the variance are squared and that it is difficult to interpret the meaning of squared units. Because the standard deviation is the square root of the variance, the units of the variance, $(\text{students})^2$ in our example, are converted to students in the standard deviation. In other words, the standard deviation is measured in the same units as the original data. For this reason, the standard deviation is more easily compared to the mean and other statistics that are measured in the same units as the original data.

Figure 2.21 shows the Excel calculation for the sample standard deviation of the home sales data, which can be calculated using Excel's STDEV.S function. The sample standard deviation in cell E9 is calculated using the formula *=STDEV.S(B2:B13)*. Excel calculates the sample standard deviation for the home sales to be $95,065.77.

Coefficient of Variation

In some situations we may be interested in a descriptive statistic that indicates how large the standard deviation is relative to the mean. This measure is called the **coefficient of variation** and is usually expressed as a percentage.

Coefficient of Variation

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\% \qquad \textbf{(2.6)}$$

Recall the class size data set used previously where the sample of five classes had sizes of

32 42 46 46 54

The sample mean class size for these data is 44 and the sample standard deviation is 8. Therefore, thew coefficient of variation is $(8/44 \times 100) = 18.2\%$. In words, the coefficient of variation tells us that the sample standard deviation is 18.2% of the value of the sample mean. The coefficient of variation for the home sales data is shown in Figure 2.21. It is calculated in cell E11 using the formula *=E9/E2*, which divides the standard deviation by the mean. The coefficient of variation for the home sales data is 43.22%. In general, the coefficient of variation is a useful statistic for comparing the relative variability of different variables, each with different standard deviations and different means.

2.7 Analyzing Distributions

In Section 2.4, we demonstrated how to create frequency, relative, and cumulative distributions for data sets. Distributions are very useful for interpreting and analyzing data.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A distribution describes the overall variability of the observed values of a variable. In this section we introduce additional ways of analyzing distributions.

Percentiles

A **percentile** is the value of a variable at which a specified (approximate) percentage of observations are below that value. The pth percentile tells us the point in the data where approximately p% of the observations have values less than the pth percentile; hence, approximately $(100 - p)$% of the observations have values greater than the pth percentile.

Colleges and universities frequently report admission test scores in terms of percentiles. For instance, suppose an applicant obtains a raw score of 54 on the verbal portion of an admission test. How this student performed in relation to other students taking the same test may not be readily apparent. However, if the raw score of 54 corresponds to the 70th percentile, we know that approximately 70% of the students scored lower than this individual, and approximately 30% of the students scored higher.

Several procedures can be used to compute the location of the pth percentile using sample data. All provide similar values, especially for large data sets. The procedure we show here is the procedure used by Excel's PERCENTILE.EXC function as well as several other statistical software packages.

To calculate the pth percentile for a data set containing n observations we must first arrange the data in ascending order (smallest value to largest value). The smallest value is in position 1, the next smallest value is in position 2, and so on. The location of the pth percentile, denoted by L_p, is computed using the following equation:

Location of the *p*th Percentile

$$L_p = \frac{p}{100}(n + 1) \tag{2.7}$$

Once we find the position of the value of the pth percentile, we have the information we need to calculate the pth percentile.

To illustrate the computation of the pth percentile, let us compute the 85th percentile for the home sales data in Table 2.8. We begin by arranging the sample of 12 starting salaries in ascending order.

	108,000	138,000	138,000	142,000	186,000	199,500	208,000	254,000	254,000	257,500	298,000	456,250
Position	1	2	3	4	5	6	7	8	9	10	11	12

The position of each observation in the sorted data is shown directly below its value. For instance, the smallest value (108,000) is in position 1, the next smallest value (138,000) is in position 2, and so on. Using equation (2.7) with $p = 85$ and $n = 12$, the location of the 85th percentile is

$$L_{85} = \frac{p}{100}(n + 1) = \left(\frac{85}{100}\right)(12 + 1) = 11.05$$

The interpretation of $L_{85} = 11.05$ is that the 85th percentile is 5% of the way between the value in position 11 and the value in position 12. In other words, the 85th percentile is the value in position 11 (298,000) plus 0.05 times the difference between the value in position 12 (456,250) and the value in position 11 (298,000). Thus, the 85th percentile is

$$\begin{aligned}\text{85th percentile} &= 298{,}000 + 0.05(456{,}250 - 298{,}000) = 298{,}000 + 0.05(158{,}250)\\ &= 305{,}912.50\end{aligned}$$

Therefore, \$305,912.50 represents the 85th percentile of the home sales data.

The pth percentile can also be calculated in Excel using the function PERCENTILE.EXC. Figure 2.21 shows the Excel calculation for the 85th percentile of the home sales data. The value in cell E13 is calculated using the formula *=PERCENTILE.EXC(B2:B13,0.85)*; B2:B13 defines the data set for which we are calculating a percentile, and 0.85 defines the percentile of interest.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Similar to percentiles, there are multiple methods for computing quartiles that all give similar results. Here we describe a commonly used method that is equivalent to Excel's QUARTILE.EXC function.

Quartiles

It is often desirable to divide data into four parts, with each part containing approximately one-fourth, or 25 percent, of the observations. These division points are referred to as the **quartiles** and are defined as follows:

Q_1 = first quartile, or 25th percentile
Q_2 = second quartile, or 50th percentile (also the median)
Q_3 = third quartile, or 75th percentile

To demonstrate quartiles, the home sales data are again arranged in ascending order.

	108,000	138,000	138,000	142,000	186,000	199,500	208,000	254,000	254,000	257,500	298,000	456,250
Position	1	2	3	4	5	6	7	8	9	10	11	12

We already identified Q_2, the second quartile (median), as 203,750. To find Q_1 and Q_3 we must find the 25th and 75th percentiles.

For Q_1,

$$L_{25} = \frac{p}{100}(n+1) = \left(\frac{25}{100}\right)(12+1) = 3.25$$

$$\begin{aligned}25\text{th percentile} &= 138{,}000 + 0.25(142{,}000 - 138{,}000) = 138{,}000 + 0.25(4{,}000)\\ &= 139{,}000\end{aligned}$$

For Q_3,

$$L_{75} = \frac{p}{100}(n+1) = \left(\frac{75}{100}\right)(12+1) = 9.75$$

$$\begin{aligned}75\text{th percentile} &= 254{,}000 + 0.75(257{,}500 - 254{,}000) = 254{,}000 + 0.75(3{,}500)\\ &= 256{,}625\end{aligned}$$

Therefore, the 25th percentile for the home sales data is $139,000 and the 75th percentile is $256,625.

The quartiles divide the home sales data into four parts, with each part containing 25% of the observations.

108,000	142,000	208,000	257,500
138,000	186,000	254,000	298,000
138,000	199,500	254,000	456,250

$Q_1 = 139{,}000$ $Q_2 = 203{,}750$ $Q_3 = 256{,}625$

The difference between the third and first quartiles is often referred to as the **interquartile range**, or IQR. For the home sales data, IQR $= Q_3 - Q_1 = 256{,}625 - 139{,}000 = 117{,}625$. Because it excludes the smallest and largest 25% of values in the data, the IQR is a useful measure of variation for data that have extreme values or are highly skewed.

A quartile can be computed in Excel using the function QUARTILE.EXC. Figure 2.21 shows the calculations for first, second, and third quartiles for the home sales data. The formula used in cell E15 is *=QUARTILE.EXC(B2:B13,1)*. The range B2:B13 defines the data set, and 1 indicates that we want to compute the first quartile. Cells E16 and E17 use similar formulas to compute the second and third quartiles.

z-Scores

A ***z*-score** allows us to measure the relative location of a value in the data set. More specifically, a *z*-score helps us determine how far a particular value is from the mean relative

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to the data set's standard deviation. Suppose we have a sample of n observations, with the values denoted by $x_1, x_2, \ldots, x_n$. In addition, assume that the sample mean, $\bar{x}$, and the sample standard deviation, s, are already computed. Associated with each value, x_i, is another value called its z-score. Equation (2.8) shows how the z-score is computed for each x_i:

z-Score

$$z_i = \frac{x_i - \bar{x}}{s} \tag{2.8}$$

where

z_i = the z-score for x_i
$\bar{x}$ = the sample mean
s = the sample standard deviation

The z-score is often called the *standardized value*. The z-score, z_i, can be interpreted as the number of standard deviations, x_i, from the mean. For example, $z_1 = 1.2$ indicates that x_1 is 1.2 standard deviations greater than the sample mean. Similarly, $z_2 = -0.5$ indicates that x_2 is 0.5, or 1/2, standard deviation less than the sample mean. A z-score greater than zero occurs for observations with a value greater than the mean, and a z-score less than zero occurs for observations with a value less than the mean. A z-score of zero indicates that the value of the observation is equal to the mean.

The z-scores for the class size data are computed in Table 2.12. Recall the previously computed sample mean, $\bar{x} = 44$, and sample standard deviation, $s = 8$. The z-score of -1.50 for the fifth observation shows that it is farthest from the mean; it is 1.50 standard deviations below the mean.

The z-score can be calculated in Excel using the function STANDARDIZE. Figure 2.22 demonstrates the use of the STANDARDIZE function to compute z-scores for the home sales data. To calculate the z-scores, we must provide the mean and standard deviation for the data set in the arguments of the STANDARDIZE function. For instance, the z-score in cell C2 is calculated with the formula *=STANDARDIZE(B2, $B$15, $B$16),* where cell B15 contains the mean of the home sales data and cell B16 contains the standard deviation of the home sales data. We can then copy and paste this formula into cells C3:C13.

Empirical Rule

When the distribution of data exhibits a symmetric bell-shaped distribution, as shown in Figure 2.23, the **empirical rule** can be used to determine the percentage of data values that are within a specified number of standard deviations of the mean. Many, but not all, distributions of data found in practice exhibit a symmetric bell-shaped distribution.

Table 2.12 z-Scores for the Class Size Data

No. of Students in Class (x_i)	Deviation About the Mean ($x_i - \bar{x}$)	z-Score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	−2	−2/8 = −0.25
46	2	2/8 = 0.25
32	−12	−12/8 = −1.50

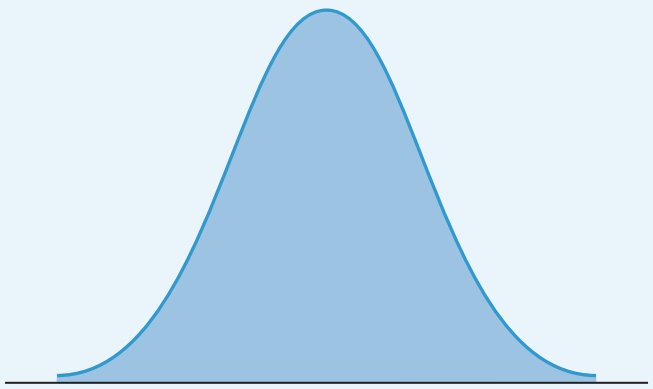
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.22 Calculating *z*-Scores for the Home Sales Data in Excel

	A	B	C
1	**Home Sale**	**Selling Price ($)**	***z*-Score**
2	1	138000	=STANDARDIZE(B2,$B$15,$B$16)
3	2	254000	=STANDARDIZE(B3,$B$15,$B$16)
4	3	186000	=STANDARDIZE(B4,$B$15,$B$16)
5	4	257500	=STANDARDIZE(B5,$B$15,$B$16)
6	5	108000	=STANDARDIZE(B6,$B$15,$B$16)
7	6	254000	=STANDARDIZE(B7,$B$15,$B$16)
8	7	138000	=STANDARDIZE(B8,$B$15,$B$16)
9	8	298000	=STANDARDIZE(B9,$B$15,$B$16)
10	9	199500	=STANDARDIZE(B10,$B$15,$B$16)
11	10	208000	=STANDARDIZE(B11,$B$15,$B$16)
12	11	142000	=STANDARDIZE(B12,$B$15,$B$16)
13	12	456250	=STANDARDIZE(B13,$B$15,$B$16)
14			
15	**Mean:**	=AVERAGE(B2:B13)	
16	**Standard Deviation:**	=STDEV.S(B2:B13)	

	A	B	C
1	**Home Sale**	**Selling Price ($)**	***z*-Score**
2	1	138,000	-0.862
3	2	254,000	0.358
4	3	186,000	-0.357
5	4	257,500	0.395
6	5	108,000	-1.177
7	6	254,000	0.358
8	7	138,000	-0.862
9	8	298,000	0.821
10	9	199,500	-0.215
11	10	208,000	-0.126
12	11	142,000	-0.820
13	12	456,250	2.486
14			
15	**Mean:**	$ 219,937.50	
16	**Standard Deviation:**	$ 95,065.77	

Figure 2.23 A Symmetric Bell-Shaped Distribution



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Empirical Rule

For data having a bell-shaped distribution:

- Approximately 68% of the data values will be within 1 standard deviation of the mean.
- Approximately 95% of the data values will be within 2 standard deviations of the mean.
- Almost all of the data values will be within 3 standard deviations of the mean.

The height of adult males in the United States has a bell-shaped distribution similar to that shown in Figure 2.23, with a mean of approximately 69.5 inches and standard deviation of approximately 3 inches. Using the empirical rule, we can draw the following conclusions.

- Approximately 68% of adult males in the United States have heights between $69.5 - 3 = 66.5$ and $69.5 + 3 = 72.5$ inches.
- Approximately 95% of adult males in the United States have heights between 63.5 and 75.5 inches.
- Almost all adult males in the United States have heights between 60.5 and 78.5 inches.

Identifying Outliers

Sometimes a data set will have one or more observations with unusually large or unusually small values. These extreme values are called **outliers**. Experienced statisticians take steps to identify outliers and then review each one carefully. An outlier may be a data value that has been incorrectly recorded; if so, it can be corrected before the data are analyzed further. An outlier may also be from an observation that doesn't belong to the population we are studying and was incorrectly included in the data set; if so, it can be removed. Finally, an outlier may be an unusual data value that has been recorded correctly and is a member of the population we are studying. In such cases, the observation should remain.

Standardized values (z-scores) can be used to identify outliers. Recall that the empirical rule allows us to conclude that for data with a bell-shaped distribution, almost all the data values will be within 3 standard deviations of the mean. Hence, in using z-scores to identify outliers, we recommend treating any data value with a z-score less than -3 or greater than $+3$ as an outlier. Such data values can then be reviewed to determine their accuracy and whether they belong in the data set.

Boxplots

A **boxplot** is a graphical summary of the distribution of data. A boxplot is developed from the quartiles for a data set. Figure 2.24 is a boxplot for the home sales data. Here are the steps used to construct the boxplot:

Boxplots are also known as box-and-whisker plots.

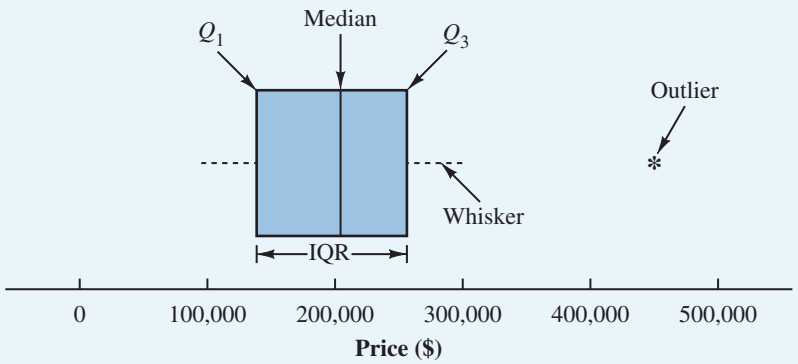
1. A box is drawn with the ends of the box located at the first and third quartiles. For the home sales data, $Q_1 = 139{,}000$ and $Q_3 = 256{,}625$. This box contains the middle 50% of the data.
2. A vertical line is drawn in the box at the location of the median (203,750 for the home sales data).

Clearly, we would not expect a home sales price less than 0, so we could also define the lower limit here to be $0.

3. By using the interquartile range, $\text{IQR} = Q_3 - Q_1$, limits are located. The limits for the boxplot are 1.5(IQR) below Q_1 and 1.5(IQR) above Q_3. For the home sales data, $\text{IQR} = Q_3 - Q_1 = 256{,}625 - 139{,}000 = 117{,}625$. Thus, the limits are $139{,}000 - 1.5(117{,}625) = -37{,}437.5$ and $256{,}625 + 1.5(117{,}625) = 433{,}062.5$. Data outside these limits are considered outliers.
4. The dashed lines in Figure 2.24 are called *whiskers*. The whiskers are drawn from the ends of the box to the smallest and largest values inside the limits computed in Step 3. Thus, the whiskers end at home sales values of 108,000 and 298,000.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.24 Boxplot for the Home Sales Data



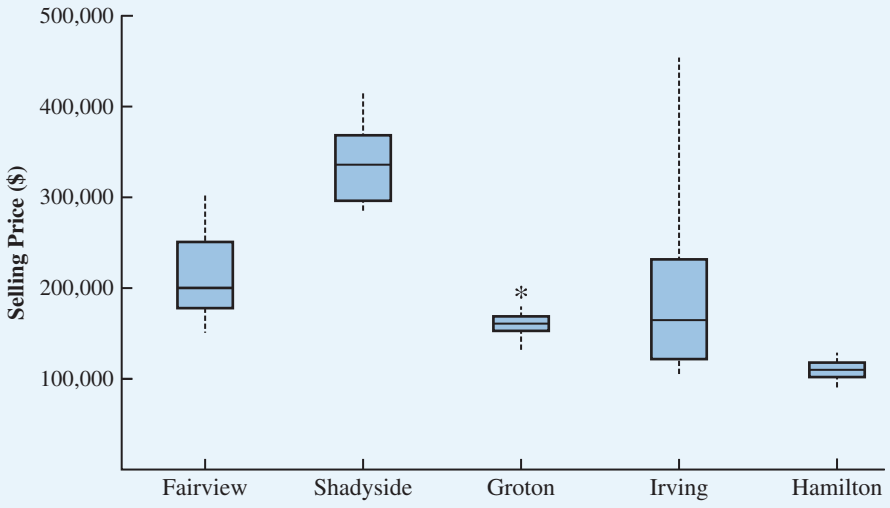
5. Finally, the location of each outlier is shown with an asterisk (*). In Figure 2.24, we see one outlier, 456,250.

Boxplots are also very useful for comparing different data sets. For instance, if we want to compare home sales from several different communities, we could create boxplots for recent home sales in each community. An example of such boxplots is shown in Figure 2.25.

What can we learn from these boxplots? The most expensive houses appear to be in Shadyside and the cheapest houses in Hamilton. The median home sales price in Groton is about the same as the median home sales price in Irving. However, home sales prices in Irving have much greater variability. Homes appear to be selling in Irving for many different prices, from very low to very high. Home sales prices have the least variation in Groton and Hamilton. The only outlier that appears in these boxplots is for home sales in Groton. However, note that most homes sell for very similar prices in Groton, so the selling price does not have to be too far from the median to be considered an outlier.

Boxplots can be drawn horizontally or vertically. Figure 2.24 shows a horizontal boxplot, and Figure 2.25 shows vertical boxplots.

Figure 2.25 Boxplots Comparing Home Sales Prices in Different Communities



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note that boxplots use a different definition of an outlier than what we described for using z-scores because the distribution of the data in a boxplot is not assumed to follow a bell-shaped curve. However, the interpretation is the same. The outliers in a boxplot are extreme values that should be investigated to ensure data accuracy.

The step-by-step directions below illustrate how to create boxplots in Excel for both a single variable and multiple variables. First we will create a boxplot for a single variable using the *homesales* file.

DATA*file*
homesales

Step 1. Select cells B1:B13
Step 2. Click the **Insert** tab on the Ribbon
Click the **Insert Statistic Chart** button in the **Charts** group
Choose the **Box and Whisker** chart from the drop-down menu

The resulting boxplot created in Excel is shown in Figure 2.26. Comparing this figure to Figure 2.24, we see that all the important elements of a boxplot are generated here. Excel orients the boxplot vertically, and by default it also includes a marker for the mean.

Next we will use the *homesalescomparison* file to create boxplots in Excel for multiple variables similar to what is shown in Figure 2.27.

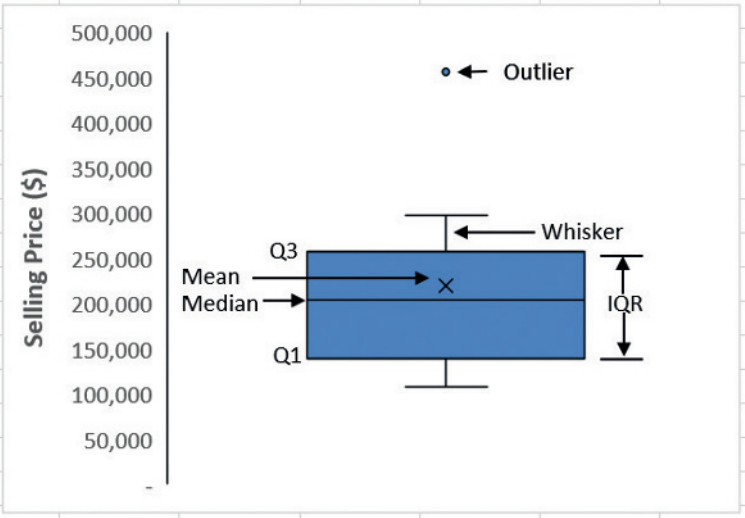
Step 1. Select cells B1:F11
Step 2. Click the **Insert** tab on the Ribbon
Click the **Insert Statistic Chart** button in the **Charts** group
Choose the **Box and Whisker** chart from the drop-down menu

DATA*file*
homesalescomparison

The boxplot created in Excel is shown in Figure 2.27. Excel again orients the boxplot vertically. The different selling locations are shown in the Legend at the top of the figure, and different colors are used for each boxplot.

Figure 2.26 Boxplot Created in Excel for Home Sales Data

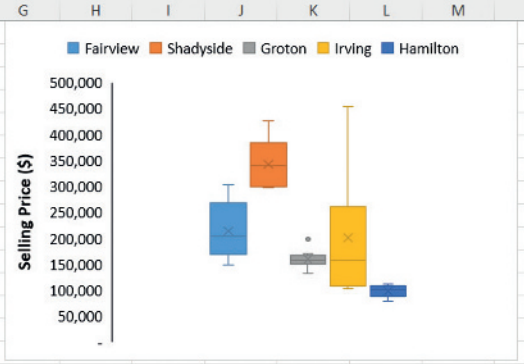
	A	B
1	Home Sale	Selling Price ($)
2	1	138,000
3	2	254,000
4	3	186,000
5	4	257,500
6	5	108,000
7	6	254,000
8	7	138,000
9	8	298,000
10	9	199,500
11	10	208,000
12	11	142,000
13	12	456,250
14		
15		
16		



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.27 Boxplots for Multiple Variables Created in Excel

	A	B	C	D	E	F
1		Fairview	Shadyside	Groton	Irving	Hamilton
2		302,000	336,000	152,000	201,000	102,000
3		265,000	398,000	158,000	365,000	108,000
4		280,000	378,000	154,000	115,000	88,000
5		220,000	298,000	170,000	105,000	111,000
6	Selling	149,000	425,000	132,000	225,000	105,000
7	Price ($)	155,000	344,000	164,000	115,000	87,000
8		198,000	302,000	198,000	108,000	95,000
9		187,000	300,000	158,000	218,000	111,000
10		208,000	298,000	149,000	454,000	98,000
11		174,000	342,000	165,000	103,000	78,000
12						
13						
14						
15						



Notes + Comments

1. The empirical rule applies only to distributions that have an approximately bell-shaped distribution because it is based on properties of the normal probability distribution, which we will discuss in Chapter 5. For distributions that do not have a bell-shaped distribution, one can use Chebyshev's theorem to make statements about the proportion of data values that must be within a specified number of standard deviations of the mean. Chebyshev's theorem states that at least $\left(1 - \frac{1}{z^2}\right) \times 100\%$ of the data values must be within z standard deviations of the mean, where z is any value greater than 1.
2. Note that the boxplot in Figure 2.26 has been formatted using Excel's Chart Elements button. These options are discussed in more detail in Chapter 3. We have also added the text descriptions of the different elements of the boxplot.

2.8 Measures of Association Between Two Variables

Thus far, we have examined numerical methods used to summarize the data for one variable at a time. Often a manager or decision maker is interested in the relationship between two variables. In this section, we present covariance and correlation as descriptive measures of the relationship between two variables. To illustrate these concepts, we consider the case of the sales manager of Queensland Amusement Park, who is in charge of ordering bottled water to be purchased by park customers. The sales manager believes that daily bottled water sales in the summer are related to the outdoor temperature. Table 2.13 shows data for high temperatures and bottled water sales for 14 summer days. The data have been sorted by high temperature from lowest value to highest value.

Scatter Charts

A scatter chart is also known as a scatter diagram or a scatter plot.

A **scatter chart** is a useful graph for analyzing the relationship between two variables. Figure 2.28 shows a scatter chart for sales of bottled water versus the high temperature experienced on 14 consecutive days. The scatter chart in the figure suggests that higher

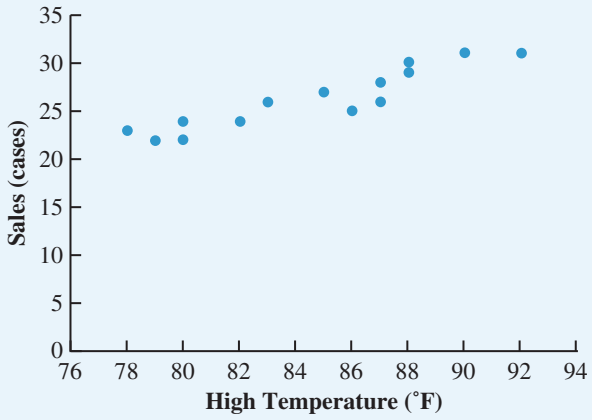
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
bottledwater

Table 2.13 Data for Bottled Water Sales at Queensland Amusement Park for a Sample of 14 Summer Days

High Temperature (°F)	Bottled Water Sales (Cases)
78	23
79	22
80	24
80	22
82	24
83	26
85	27
86	25
87	28
87	26
88	29
88	30
90	31
92	31

Figure 2.28 Chart Showing the Positive Linear Relation Between Sales and High Temperatures



daily high temperatures are associated with higher bottled water sales. This is an example of a positive relationship, because when one variable (high temperature) increases, the other variable (sales of bottled water) generally also increases. The scatter chart also suggests that a straight line could be used as an approximation for the relationship between high temperature and sales of bottled water.

Scatter charts are discussed in Chapter 3.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Covariance

Covariance is a descriptive measure of the linear association between two variables. For a sample of size n with the observations (x_1, y_1), (x_2, y_2), and so on, the sample covariance is defined as follows:

Sample Covariance

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{2.9}$$

If data consist of a population of N *observations, the population covariance* σ_{xy} *is computed by:*

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)\Sigma(y_i - \mu_y)}{N}.$$

Note that this equation is similar to equation (2.9), but uses population parameters instead of sample estimates (and divides by N *instead of* n − 1 *for technical reasons beyond the scope of this book).*

This formula pairs each x_i with a y_i. We then sum the products obtained by multiplying the deviation of each x_i from its sample mean $(x_i - \bar{x})$ by the deviation of the corresponding y_i from its sample mean $(y_i - \bar{y})$; this sum is then divided by $n - 1$.

To measure the strength of the linear relationship between the high temperature x and the sales of bottled water y at Queensland, we use equation (2.9) to compute the sample covariance. The calculations in Table 2.14 show the computation $\Sigma(x_i - \bar{x})(y_i - \bar{y})$. Note that for our calculations, $\bar{x} = 84.6$ and $\bar{y} = 26.3$.

The covariance calculated in Table 2.14 is $s_{xy} = 12.8$. Because the covariance is greater than 0, it indicates a positive relationship between the high temperature and sales of bottled water. This verifies the relationship we saw in the scatter chart in Figure 2.28 that as the high temperature for a day increases, sales of bottled water generally increase.

The sample covariance can also be calculated in Excel using the COVARIANCE.S function. Figure 2.29 shows the data from Table 2.13 entered into an Excel Worksheet. The covariance is calculated in cell B17 using the formula *=COVARIANCE.S(A2:A15, B2:B15)*.

Table 2.14 Sample Covariance Calculations for Daily High Temperature and Bottled Water Sales at Queensland Amusement Park

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	78	23	−6.6	−3.3	21.78
	79	22	−5.6	−4.3	24.08
	80	24	−4.6	−2.3	10.58
	80	22	−4.6	−4.3	19.78
	82	24	−2.6	−2.3	5.98
	83	26	−1.6	−0.3	0.48
	85	27	0.4	0.7	0.28
	86	25	1.4	−1.3	−1.82
	87	28	2.4	1.7	4.08
	87	26	2.4	−0.3	−0.72
	88	29	3.4	2.7	9.18
	88	30	3.4	3.7	12.58
	90	31	5.4	4.7	25.38
	92	31	7.4	4.7	34.78
Totals	1,185	368	0.6	−0.2	166.42

$\bar{x} = 84.6$

$\bar{y} = 26.3$

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{166.42}{14 - 1} = 12.8$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.29 Calculating Covariance and Correlation Coefficient for Bottled Water Sales Using Excel

	A	B
1	**High Temperature (degrees F)**	**Bottled Water Sales (cases)**
2	78	23
3	79	22
4	80	24
5	80	22
6	82	24
7	83	26
8	85	27
9	86	25
10	87	28
11	87	26
12	88	29
13	88	30
14	90	31
15	92	31
16		
17	**Covariance:**	=COVARIANCE.S(A2:A15,B2:B15)
18	**Correlation Coefficient:**	=CORREL(A2:A15,B2:B15)

	A	B
1	**High Temperature (degrees F)**	**Bottled Water Sales (cases)**
2	78	23
3	79	22
4	80	24
5	80	22
6	82	24
7	83	26
8	85	27
9	86	25
10	87	28
11	87	26
12	88	29
13	88	30
14	90	31
15	92	31
16		
17	**Covariance:**	12.80
18	**Correlation Coefficient:**	0.93

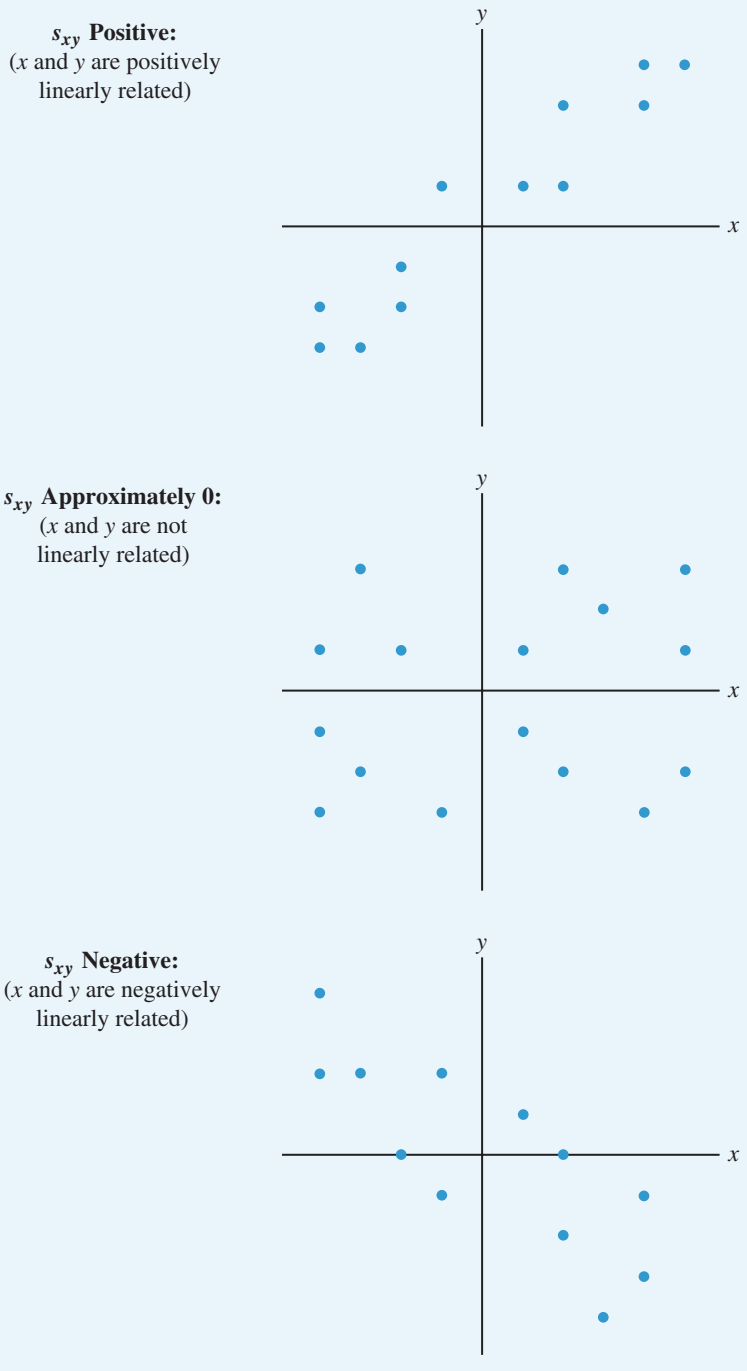
A2:A15 defines the range for the x variable (high temperature), and B2:B15 defines the range for the y variable (sales of bottled water).

For the bottled water, the covariance is positive, indicating that higher temperatures (x) are associated with higher sales (y). If the covariance is near 0, then the x and y variables are not linearly related. If the covariance is less than 0, then the x and y variables are negatively related, which means that as x increases, y generally decreases. Figure 2.30 demonstrates several possible scatter charts and their associated covariance values.

One problem with using covariance is that the magnitude of the covariance value is difficult to interpret. Larger s_{xy} values do not necessarily mean a stronger linear relationship because the units of covariance depend on the units of x and y. For example, suppose we are interested in the relationship between height x and weight y for individuals. Clearly the strength of the relationship should be the same whether we measure height in feet or inches. Measuring the height in inches, however, gives us much larger numerical values for $(x_i - \bar{x})$ than when we measure height in feet. Thus, with height measured in inches, we would obtain a larger value for the numerator $\Sigma(x_i - \bar{x})(y_i - \bar{y})$ in equation (2.9)—and hence a larger covariance—when in fact the relationship does not change.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.30 Scatter Charts and Associated Covariance Values for Different Variable Relationships



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Correlation Coefficient

The **correlation coefficient** measures the relationship between two variables, and, unlike covariance, the relationship between two variables is not affected by the units of measurement for x and y. For sample data, the correlation coefficient is defined as follows:

Sample Correlation Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{2.10}$$

where

r_{xy} = sample correlation coefficient
s_{xy} = sample covariance
s_x = sample standard deviation of x
s_y = sample standard deviation of y

If data are a population, the population correlation coefficient is computed by $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$. Note that this is similar to equation (2.10) but uses population parameters instead of sample estimates.

The sample correlation coefficient is computed by dividing the sample covariance by the product of the sample standard deviation of x and the sample standard deviation of y. This scales the correlation coefficient so that it will always take values between -1 and $+1$.

Let us now compute the sample correlation coefficient for bottled water sales at Queensland Amusement Park. Recall that we calculated $s_{xy} = 12.8$ using equation (2.9). Using data in Table 2.13, we can compute sample standard deviations for x and y.

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = 4.36$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = 3.15$$

The sample correlation coefficient is computed from equation (2.10) as follows:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{12.8}{(4.36)(3.15)} = 0.93$$

The correlation coefficient can take only values between -1 and $+1$. Correlation coefficient values near 0 indicate no linear relationship between the x and y variables. Correlation coefficients greater than 0 indicate a positive linear relationship between the x and y variables. The closer the correlation coefficient is to $+1$, the closer the x and y values are to forming a straight line that trends upward to the right (positive slope). Correlation coefficients less than 0 indicate a negative linear relationship between the x and y variables. The closer the correlation coefficient is to -1, the closer the x and y values are to forming a straight line with negative slope. Because $r_{xy} = 0.93$ for the bottled water, we know that there is a very strong positive linear relationship between these two variables. As we can see in Figure 2.28, one could draw a straight line with a positive slope that would be very close to all of the data points in the scatter chart.

Because the correlation coefficient defined here measures only the strength of the linear relationship between two quantitative variables, it is possible for the correlation coefficient to be near zero, suggesting no linear relationship, when the relationship between the two variables is nonlinear. For example, the scatter chart in Figure 2.31 shows the relationship between monthly utility bills for heating/cooling and the mean high temperature during that month.

The sample correlation coefficient for these data is $r_{xy} = 0.01$ and indicates that there is no linear relationship between the two variables. However, it would be incorrect to suggest that there is no relationship between these two variables. Figure 2.31 provides strong visual evidence of a nonlinear relationship as indicated by the curved dashed trend line in

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 2.31 Example of Nonlinear Relationship Producing a Correlation Coefficient Near Zero

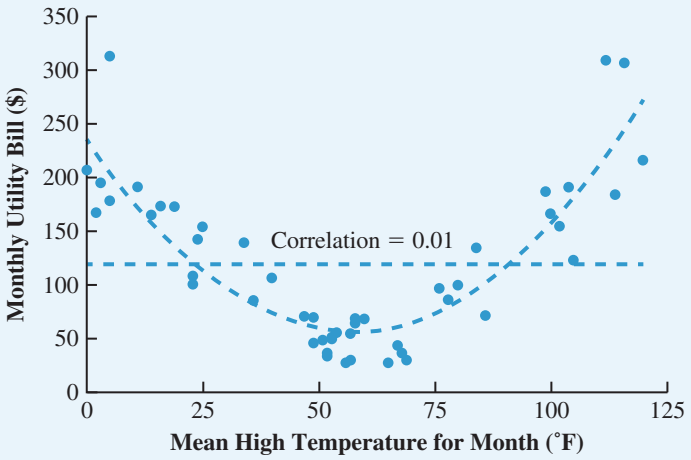


Figure 2.31. That is, we can see that as the mean high temperature is either relatively low or relatively high, the cost of the monthly utility bill increases.

We can compute correlation coefficients using the Excel function CORREL. The correlation coefficient in Figure 2.29 is computed in cell B18 for the sales of bottled water using the formula *=CORREL(A2:A15, B2:B15)*, where A2:A15 defines the range for the *x* variable and B2:B15 defines the range for the *y* variable.

Notes + Comments

1. The correlation coefficient discussed in this chapter was developed by Karl Pearson and is sometimes referred to as Pearson product moment correlation coefficient. It is appropriate for use only with two quantitative variables. A variety of alternatives, such as the Spearman rank-correlation coefficient, exist to measure the association of categorical variables. The Spearman rank-correlation coefficient is discussed in Chapter 13.
2. Correlation measures only the association between two variables. A large positive or large negative correlation coefficient does not indicate that a change in the value of one of the two variables *causes* a change in the value of the other variable.

Summary

In this chapter, we have provided an introduction to descriptive statistics that can be used to summarize data. We began by explaining the need for data collection, defining the types of data one may encounter, and providing a few commonly used sources for finding data. We presented several useful functions for modifying data in Excel, such as sorting and filtering to aid in data analysis.

We introduced the concept of a distribution and explained how to generate frequency, relative, percent, and cumulative distributions for data. We also demonstrated the use of histograms and frequency polygons to visualize the distribution of data. We then

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

introduced measures of location for a distribution of data such as mean, median, mode, and geometric mean, as well as measures of variability such as range, variance, standard deviation, coefficient of variation, and interquartile range. We presented additional measures for analyzing a distribution of data including percentiles, quartiles, and z-scores. We showed that boxplots are effective for visualizing a distribution.

We discussed measures of association between two variables. Scatter plots allow one to visualize the relationship between variables. Covariance and the correlation coefficient summarize the linear relationship between variables into a single number. We showed that the correlation coefficient only measures linear relationships between two variables and not nonlinear relationships.

Glossary

Bins The nonoverlapping groupings of data used to create a frequency distribution. Bins for categorical data are also known as classes.

Boxplot A graphical summary of data based on the quartiles of a distribution.

Categorical data Data for which categories of like items are identified by labels or names. Arithmetic operations cannot be performed on categorical data.

Coefficient of variation A measure of relative variability computed by dividing the standard deviation by the mean and multiplying by 100.

Correlation coefficient A standardized measure of linear association between two variables that takes on values between -1 and $+1$. Values near -1 indicate a strong negative linear relationship, values near $+1$ indicate a strong positive linear relationship, and values near zero indicate the lack of a linear relationship.

Covariance A measure of linear association between two variables. Positive values indicate a positive relationship; negative values indicate a negative relationship.

Cross-sectional data Data collected at the same or approximately the same point in time.

Cumulative frequency distribution A tabular summary of quantitative data showing the number of data values that are less than or equal to the upper limit of each bin.

Data The facts and figures collected, analyzed, and summarized for presentation and interpretation.

Empirical rule A rule that can be used to compute the percentage of data values that must be within 1, 2, or 3 standard deviations of the mean for data that exhibit a bell-shaped distribution.

Frequency distribution A tabular summary of data showing the number (frequency) of data values in each of several nonoverlapping bins.

Frequency polygon A chart used to display a distribution by using lines to connect the frequency values of each bin.

Geometric mean A measure of central location that is calculated by finding the nth root of the product of n values.

Growth factor The percentage increase of a value over a period of time is calculated using the formula (growth factor $-$ 1). A growth factor less than 1 indicates negative growth, whereas a growth factor greater than 1 indicates positive growth. The growth factor cannot be less than zero.

Histogram A graphical presentation of a frequency distribution, relative frequency distribution, or percent frequency distribution of quantitative data constructed by placing the bin intervals on the horizontal axis and the frequencies, relative frequencies, or percent frequencies on the vertical axis.

Interquartile range The difference between the third and first quartiles.

Mean (arithmetic mean) A measure of central location computed by summing the data values and dividing by the number of observations.

Median A measure of central location provided by the value in the middle when the data are arranged in ascending order.

Mode A measure of central location defined as the value that occurs with greatest frequency.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Observation A set of values corresponding to a set of variables.
Outliers An unusually large or unusually small data value.
Percent frequency distribution A tabular summary of data showing the percentage of data values in each of several nonoverlapping bins.
Percentile A value such that approximately $p\%$ of the observations have values less than the pth percentile; hence, approximately $(100 - p)\%$ of the observations have values greater than the pth percentile. The 50th percentile is the median.
Population The set of all elements of interest in a particular study.
Quantitative data Data for which numerical values are used to indicate magnitude, such as how many or how much. Arithmetic operations such as addition, subtraction, and multiplication can be performed on quantitative data.
Quartiles The 25th, 50th, and 75th percentiles, referred to as the first quartile, second quartile (median), and third quartile, respectively. The quartiles can be used to divide a data set into four parts, with each part containing approximately 25% of the data.
Random sampling Collecting a sample that ensures that (1) each element selected comes from the same population and (2) each element is selected independently.
Random variable, or uncertain variable A quantity whose values are not known with certainty.
Range A measure of variability defined to be the largest value minus the smallest value.
Relative frequency distribution A tabular summary of data showing the fraction or proportion of data values in each of several nonoverlapping bins.
Sample A subset of the population.
Scatter chart A graphical presentation of the relationship between two quantitative variables. One variable is shown on the horizontal axis and the other on the vertical axis.
Skewness A measure of the lack of symmetry in a distribution.
Standard deviation A measure of variability computed by taking the positive square root of the variance.
Time series data Data that are collected over a period of time (minutes, hours, days, months, years, etc.).
Variable A characteristic or quantity of interest that can take on different values.
Variance A measure of variability based on the squared deviations of the data values about the mean.
Variation Differences in values of a variable over observations.
***z*-score** A value computed by dividing the deviation about the mean $(x_i - x)$ by the standard deviation s. A z-score is referred to as a standardized value and denotes the number of standard deviations that x_i is from the mean.

Problems

1. ***Wall Street Journal* Subscriber Characteristics.** A *Wall Street Journal* subscriber survey asked 46 questions about subscriber characteristics and interests. State whether each of the following questions provides categorical or quantitative data. **LO 1**
   a. What is your age?
   b. Are you male or female?
   c. When did you first start reading the *WSJ*? High school, college, early career, midcareer, late career, or retirement?
   d. How long have you been in your present job or position?
   e. What type of vehicle are you considering for your next purchase? Nine response categories include sedan, sports car, SUV, minivan, and so on.
2. **Gross Domestic Products.** The following table contains a partial list of countries, the continents on which they are located, and their respective gross domestic products (GDPs) in U.S. dollars. A list of 125 countries and their GDPs is contained in the file *gdplist*. **LO 2**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
gdplist

Country	Continent	GDP (US$ millions)
Afghanistan	Asia	18,181
Albania	Europe	12,847
Algeria	Africa	190,709
Angola	Africa	100,948
Argentina	South America	447,644
Australia	Oceania	1,488,221
Austria	Europe	419,243
Azerbaijan	Europe	62,321
Bahrain	Asia	26,108
Bangladesh	Asia	113,032
Belarus	Europe	55,483
Belgium	Europe	513,396
Bolivia	South America	24,604
Bosnia and Herzegovina	Europe	17,965
Botswana	Africa	17,570

a. Sort the countries in the file *gdplist* from largest to smallest GDP. What are the top 10 countries according to GDP?
b. Filter the countries to display only the countries located in Africa. What are the top five countries located in Africa according to GDP?
c. What are the top five countries by GDP that are located in Europe?

3. **On-Time Performance of Logistics Companies.** Ohio Logistics manages the logistical activities for firms by matching companies that need products shipped with carriers that can provide the best rates and best service for the companies. Ohio Logistics is very concerned that its carriers deliver their customers' material on time, so it carefully monitors the percentage of on-time deliveries. The following table contains a list of the carriers used by Ohio Logistics and the corresponding on-time percentages for the current and previous years. **LO 2**

DATA *file*
carriers

Carrier	Previous Year On-Time Deliveries (%)	Current Year On-Time Deliveries (%)
Blue Box Shipping	88.4	94.8
Cheetah LLC	89.3	91.8
Granite State Carriers	81.8	87.6
Honsin Limited	74.2	80.1
Jones Brothers	68.9	82.8
Minuteman Company	91.0	84.2
Rapid Response	78.8	70.9
Smith Logistics	84.3	88.7
Super Freight	92.1	86.8

a. Sort the carriers in descending order by their current year's percentage of on-time deliveries. Which carrier is providing the best service in the current year? Which carrier is providing the worst service in the current year?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Calculate the change in percentage of on-time deliveries from the previous to the current year for each carrier. Use Excel's conditional formatting to highlight the carriers whose on-time percentage decreased from the previous year to the current year.
c. Use Excel's conditional formatting tool to create data bars for the change in percentage of on-time deliveries from the previous year to the current year for each carrier calculated in part b.
d. Which carriers should Ohio Logistics try to use in the future? Why?

4. **Relative Frequency Distribution.** A partial relative frequency distribution is given. LO 3

Class	Relative Frequency
A	0.22
B	0.18
C	0.40
D	

a. What is the relative frequency of class D?
b. The total sample size is 200. What is the frequency of class D?
c. Show the frequency distribution.
d. Show the percent frequency distribution.

5. **Most Visited Web Sites.** In a recent report, the top five most-visited English-language web sites were google.com (GOOG), facebook.com (FB), youtube.com (YT), yahoo.com (YAH), and wikipedia.com (WIKI). The most-visited web sites for a sample of 50 Internet users are shown in the following table. LO 1, 3

DATA*file* websites

YAH	WIKI	YT	WIKI	GOOG
YT	YAH	GOOG	GOOG	GOOG
WIKI	GOOG	YAH	YAH	YAH
YAH	YT	GOOG	YT	YAH
GOOG	FB	FB	WIKI	GOOG
GOOG	GOOG	FB	FB	WIKI
FB	YAH	YT	YAH	YAH
YT	GOOG	YAH	FB	FB
WIKI	GOOG	YAH	WIKI	WIKI
YAH	YT	GOOG	GOOG	WIKI

a. Are these data categorical or quantitative?
b. Provide frequency and percent frequency distributions.
c. On the basis of the sample, which web site is most frequently the most-often-visited web site for Internet users? Which is second?

6. **CEO Time in Meetings.** In a study of how chief executive officers (CEOs) spend their days, it was found that CEOs spend an average of about 18 hours per week in meetings, not including conference calls, business meals, and public events. Shown here are the times spent per week in meetings (hours) for a sample of 25 CEOs follows. LO 2, 4, 5, 7

DATA*file* ceotime

14	15	18	23	15
19	20	13	15	23
23	21	15	20	21
16	15	18	18	19
19	22	23	21	12

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. What is the least amount of time a CEO spent per week in meetings in this sample? The highest?
b. Use a bin width of 2 hours to prepare a frequency distribution and a percent frequency distribution for the data.
c. Prepare a histogram and comment on the shape of the distribution.

DATA *file*
bbbcomplaints

7. **Complaints Reported to the Better Business Bureau.** Consumer complaints are frequently reported to the Better Business Bureau (BBB). Industries with the most complaints to the BBB are often banks, cable and satellite television companies, collection agencies, cellular phone providers, and new car dealerships. The results for a sample of 200 complaints are in the file *bbbcomplaints*. **LO 2**
   a. Show the frequency and percent frequency of complaints by industry.
   b. Which industry had the highest number of complaints?
   c. Comment on the percentage frequency distribution for complaints.

8. **Busiest Airports in the World.** Based on the total passenger traffic, the airports in the following list are the 25 busiest airports in the world in 2021 (Airports Council International website). **LO 2, 4, 5**

DATA *file*
worldairports

Airport	Airport Code	Location (country)	Total Passengers (millions)
Beijing Capital	PEK	China	32.6
Charlotte Douglas	CLT	United States	43.3
Chengdu Shuangliu	CTU	China	40.1
Chongqing Jiangbei	CKG	China	35.8
Dallas Fort Worth	DFW	United States	62.5
Denver	DEN	United States	58.8
Dubai	DXB	United Arab Emirates	29.1
George Bush	IAH	United States	33.7
Guangzhou Baiyun	CAN	China	40.2
Harry Reid	LAS	United States	39.7
Hartsfield–Jackson Atlanta	ATL	United States	75.5
Istanbul Airport	IST	Turkey	37.0
Kunming Changshui	KMG	China	32.2
Los Angeles	LAX	United States	48.0
Mexico City	MEX	Mexico	36.1
Miami	MIA	United States	37.3
O'Hare	ORD	United States	54.0
Orlando	MCO	United States	40.4
Phoenix Sky Harbor	PHX	United States	38.8
Seattle–Tacoma	SEA	United States	36.2
Shanghai Hongqiao	SHA	China	33.2
Shanghai Pudong	PVG	China	32.2
Shenzhen Bao'an	SZX	China	36.4
Sheremetyevo	SVO	Russia	30.9
Xi'an Xianyang	XIY	Asia	30.2

a. Which is busiest airport in the world in terms of total passenger traffic? Which is the least busy airport in terms of total passenger traffic?
b. Using a bin width of 10, develop a frequency distribution of the data using bins of [20, 30], (30, 40], (40, 50], and so on.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Create and interpret a histogram for this distribution of data.
d. Create a frequency polygon to compare the distribution of total passenger traffic for airports located in the United States to the distribution of total passenger traffic for airports located outside the United States. What can you infer about these distributions relative to one another from the frequency polygon?

9. **Vancouver Aquarium Volunteer Ages.** The minimum age to volunteer at the Vancouver Aquarium in Vancouver, British Columbia is 14 years old. The data below displays the ages of potential volunteers who have expressed interest in working at the Vancouver Aquarium. **LO 4**
   a. Develop a frequency distribution using bins of [12, 14], (14, 16], (16, 18], (18, 20], (20, 22], (22, 24], (24, 26].
   b. Develop a relative frequency distribution and a percent frequency distribution using the bins in part a.

DATA*file*
vancouveraquarium

14	24	18	22
19	18	16	22
24	17	15	16
19	23	24	16
16	26	21	16
20	22	16	14
24	23	19	25
20	25	21	19
21	25	23	24
22	19	20	20

10. **Cumulative Frequency Distribution.** Consider the following frequency distribution.

Bin	Frequency
10–19	10
20–29	14
30–39	17
40–49	7
50–59	2

Construct a cumulative frequency distribution. **LO 6**

11. **Years in the Military.** A military base in the United States is examining the distribution of the amount of time that officers on the base have been in the military. The following data show the years of military service for each officer on the base. **LO 4, 6**

DATA*file*
military_years

2 5 10 12 4 4 5 17 11 8 9 8 12 21 6 8 7 13 18 3

Using bins of [0, 5], (5, 10], (10, 15], and so on, show the following:
   a. The frequency distribution
   b. The relative frequency distribution
   c. The cumulative frequency distribution
   d. The cumulative relative frequency distribution
   e. The proportion of officers who have been in the military for more than 15 years.

12. **Largest University Endowments.** University endowments are financial assets that are donated by supporters to be used to provide income to universities. There is a large discrepancy in the size of university endowments. The following table provides a listing of many of the universities that have the largest endowments as reported by the National Association of College and University Business Officers in 2017. **LO 2, 4, 5, 6, 7**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
endowments

University	Endowment Amount ($ billions)	University	Endowment Amount ($ billions)
Amherst College	2.2	Smith College	1.8
Boston College	2.3	Stanford University	24.8
Boston University	2.0	Swarthmore College	2.0
Brown University	3.2	Texas A&M University	11.6
California Institute of Technology	2.6	Tufts University	1.7
Carnegie Mellon University	2.2	University of California	9.8
Case Western Reserve University	1.8	University of California, Berkeley	1.8
Columbia University	10.0	University of California, Los Angeles	2.1
Cornell University	6.8	University of Chicago	7.5
Dartmouth College	5.0	University of Illinois	2.6
Duke University	7.9	University of Michigan	10.9
Emory University	6.9	University of Minnesota	3.5
George Washington University	1.7	University of North Carolina at Chapel Hill	3.0
Georgetown University	1.7	University of Notre Dame	9.4
Georgia Institute of Technology	2.0	University of Oklahoma	1.6
Grinnell College	1.9	University of Pennsylvania	12.2
Harvard University	36.0	University of Pittsburgh	3.9
Indiana University	2.2	University of Richmond	2.4
Johns Hopkins University	3.8	University of Rochester	2.1
Massachusetts Institute of Technology	15.0	University of Southern California	5.1
Michigan State University	2.7	University of Texas	26.5
New York University	4.0	University of Virginia	8.6
Northwestern University	10.4	University of Washington	2.5
Ohio State University	4.3	University of Wisconsin–Madison	2.7
Pennsylvania State University	4.0	Vanderbilt University	4.1
Pomona College	2.2	Virginia Commonwealth University	1.8
Princeton University	23.8	Washington University in St. Louis	7.9
Purdue University	2.4	Wellesley College	1.9
Rice University	5.8	Williams College	2.5
Rockefeller University	2.0	Yale University	27.2

Summarize the data by constructing the following:

a. A frequency distribution using bins of [0, 2], (2, 4], (4, 6], (6, 8], (8, 10], and so on.
b. A relative frequency distribution.
c. A cumulative frequency distribution.
d. A cumulative relative frequency distribution.
e. What do these distributions tell you about the endowments of universities?
f. Show a histogram. Comment on the shape of the distribution.
g. What is the largest university endowment and which university holds it?

13. **Computing Mean and Median.** Consider a sample with data values of 10, 20, 12, 17, and 16. LO 8

a. Compute the mean and median.
b. Consider a sample with data values 10, 20, 12, 17, 16, and 12. How would you expect the mean and median for these sample data to compare to the mean and

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

median for part a (higher, lower, or the same)? Compute the mean and median for the sample data 10, 20, 12, 17, 16, and 12.

14. **Federal Tax Payment Errors.** A sample of 10,001 federal income tax returns with errors has been collected by the Internal Revenue Service (IRS) for analysis. A portion of the data is shown in the table below, and the full data set is in the file *fedtax_errors*. Each tax payment error is noted to be either an underpayment or an overpayment. Underpayments are also denoted as positive tax payment error values and overpayments as negative tax payment error values. **LO 2, 5, 7, 8, 9**

DATA*file* fedtax_errors

Tax Payment Error ($)	Underpayment?
12,616	Yes
14,231	Yes
2,043	No
4,437	No
917	Yes
11,626	No
9,215	No
821	Yes
⋮	⋮

a. Sort the data by the Underpayment? variable to group the underpayment values separately from the overpayment values.
b. Calculate the mean and standard deviation for tax payment errors that are overpayments.
c. Calculate the mean and standard deviation for tax payment errors that are underpayments.
d. Create a frequency polygon to compare the distribution of federal tax overpayments to the distribution of federal tax underpayments. Use bin widths of $2000 and start with $0. Describe any differences you see in these distributions based on the frequency polygon.

15. **Computing Percentiles.** Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28, and 25. Compute the 20th, 25th, 65th, and 75th percentiles. **LO 10**

16. **Computing Mean, Median, and Mode.** Consider a sample with data values of 53, 55, 70, 58, 64, 57, 53, 69, 57, 68, and 53. Compute the mean, median, and mode. **LO 8**

17. **Mean Annual Growth Rate of Asset.** If an asset declines in value from $5,000 to $3,500 over nine years, what is the mean annual growth rate in the asset's value over these nine years? **LO 8**

18. **Comparing Mutual Fund Investments.** Suppose that you initially invested $10,000 in the Stivers mutual fund and $5,000 in the Trippi mutual fund. The value of each investment at the end of each subsequent year is provided in the table that follows.

DATA*file* stiverstrippi

Year	Stivers ($)	Trippi ($)
1	11,000	5,600
2	12,000	6,300
3	13,000	6,900
4	14,000	7,600
5	15,000	8,500
6	16,000	9,200
7	17,000	9,900
8	18,000	10,600

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Which of the two mutual funds performed better over this time period? **LO 8**

19. **Commute Times.** The average time that Americans commute to work is 27.7 minutes (*Sterling's Best Places*). The average commute times in minutes for 48 cities follows. **LO 8, 9, 10**

DATA *file*
commutetimes

City	Commute Time (min)	City	Commute Time (min)	City	Commute Time (min)
Albuquerque	23.3	Kansas City	23.4	Portland	26.4
Atlanta	28.3	Las Vegas	28.4	Providence	23.6
Austin	24.6	Little Rock	20.1	Richmond	23.4
Baltimore	32.1	Los Angeles	32.2	Sacramento	25.8
Boston	31.7	Louisville	21.4	Salt Lake City	20.2
Charlotte	25.8	Memphis	23.8	San Antonio	26.1
Chicago	38.1	Miami	30.7	San Diego	24.8
Cincinnati	24.9	Milwaukee	24.8	San Francisco	32.6
Cleveland	26.8	Minneapolis	23.6	San Jose	28.5
Columbus	23.4	Nashville	25.3	Seattle	27.3
Dallas	28.5	New Orleans	31.7	St. Louis	26.8
Denver	28.1	New York	43.8	Tucson	24.0
Detroit	29.3	Oklahoma City	22.0	Tulsa	20.1
El Paso	24.4	Orlando	27.1	Washington, D.C.	32.8
Fresno	23.0	Philadelphia	34.2		
Indianapolis	24.8	Phoenix	28.3		
Jacksonville	26.2	Pittsburgh	25.0		

a. What is the mean commute time for these 48 cities?
b. What is the median commute time for these 48 cities?
c. What is the mode for these 48 cities?
d. What is the variance and standard deviation of commute times for these 48 cities?
e. What is the third quartile of commute times for these 48 cities?

20. **Patient Waiting Times.** Suppose that the average waiting time for a patient at a physician's office is just over 29 minutes. To address the issue of long patient wait times, some physicians' offices are using wait-tracking systems to notify patients of expected wait times. Patients can adjust their arrival times based on this information and spend less time in waiting rooms. The following data show wait times (in minutes) for a sample of patients at offices that do not have a wait-tracking system and wait times for a sample of patients at offices with such systems. **LO 8, 9, 12**

DATA *file*
patientwaits

Without Wait-Tracking System	With Wait-Tracking System
24	31
67	11
17	14
20	18
31	12
44	37
12	9
23	13
16	12
37	15

a. What are the mean and median patient wait times for offices with a wait-tracking system? What are the mean and median patient wait times for offices without a wait-tracking system?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. What are the variance and standard deviation of patient wait times for offices with a wait-tracking system? What are the variance and standard deviation of patient wait times for visits to offices without a wait-tracking system?
c. Create a boxplot for patient wait times for offices without a wait-tracking system.
d. Create a boxplot for patient wait times for offices with a wait-tracking system.
e. Do offices with a wait-tracking system have shorter patient wait times than offices without a wait-tracking system? Explain.

21. **Number of Hours Worked per Week by Teachers.** According to the National Education Association (NEA), teachers generally spend more than 40 hours each week working on instructional duties. The following data show the number of hours worked per week for a sample of 13 high school science teachers and a sample of 11 high school English teachers. **LO 8, 12**

DATA*file* teachers

High school science teachers	53	56	54	54	55	58	49	61	54	54	52	53	54
High school English teachers	52	47	50	46	47	48	49	46	55	44	47		

a. What is the median number of hours worked per week for the sample of 13 high school science teachers?
b. What is the median number of hours worked per week for the sample of 11 high school English teachers?
c. Create a boxplot for the number of hours worked for high school science teachers.
d. Create a boxplot for the number of hours worked for high school English teachers.
e. Comment on the differences between the boxplots for science and English teachers.

DATA*file* primetime_xfl

22. **Television Viewership of Football Games.** Companies considering purchasing ads during televised XFL football games in the United States are interested in the viewership of games as a way of estimating the exposure their ads get. The file *primetime_xfl* contains a sample of viewership numbers for XFL games televised on Friday and Saturday nights, respectively. **LO 10, 12**
a. Construct boxplots to compare the viewership distribution of Friday night XFL games to the viewership distribution of Saturday night XFL games.
b. Describe the distribution of the viewership of Friday night XFL games based on the boxplot.
c. Describe the distribution of the viewership of Saturday night XFL games based on the boxplot.
d. Compare the two distributions to each other based on the boxplots. Which night appears to have more viewers?

DATA*file* patientwaits

23. ***z*-Scores for Patient Waiting Times.** Return to the waiting times given for the physician's office in Problem 20. **LO 10, 11**
a. Considering only offices *without* a wait-tracking system, what is the z-score for the 10th patient in the sample (wait time = 37 minutes)?
b. Considering only offices *with* a wait-tracking system, what is the z-score for the 6th patient in the sample (wait time = 37 minutes)? How does this z-score compare with the z-score you calculated for part a?
c. Based on z-scores, do the data for offices without a wait-tracking system contain any outliers? Based on z-scores, do the data for offices without a wait-tracking system contain any outliers?

24. **Amount of Sleep per Night.** The results of a national survey showed that on average adults sleep 6.9 hours per night. Suppose that the standard deviation is 1.2 hours and that the number of hours of sleep follows a bell-shaped distribution. **LO 10**
a. Use the empirical rule to calculate the percentage of individuals who sleep between 4.5 and 9.3 hours per day.
b. What is the z-score for an adult who sleeps 8 hours per night?
c. What is the z-score for an adult who sleeps 6 hours per night?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

25. **GMAT Exam Scores.** The Graduate Management Admission Test (GMAT) is a standardized exam used by many universities as part of the assessment for admission to graduate study in business. The average GMAT score is 547 (Magoosh web site). Assume that GMAT scores are bell-shaped with a standard deviation of 100. **LO 10**
    a. What percentage of GMAT scores are 647 or higher?
    b. What percentage of GMAT scores are 747 or higher?
    c. What percentage of GMAT scores are between 447 and 547?
    d. What percentage of GMAT scores are between 347 and 647?

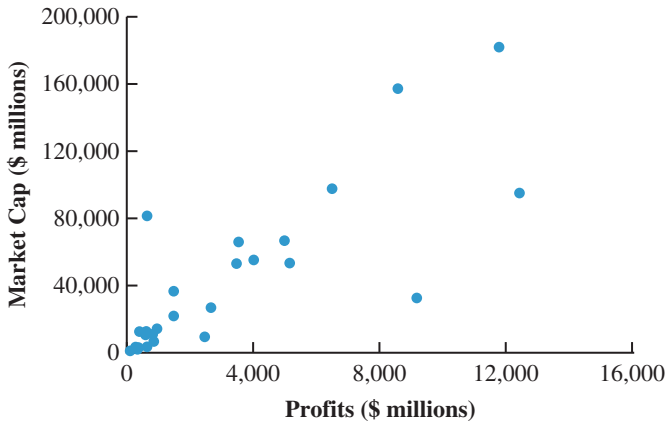
26. **Scatter Chart.** Five observations taken for two variables follow. **LO 13, 14**

x_i	4	6	11	3	16
y_i	50	50	40	60	30

    a. Develop a scatter chart with x on the horizontal axis.
    b. What does the scatter chart developed in part a indicate about the relationship between the two variables?
    c. Compute and interpret the sample covariance.
    d. Compute and interpret the sample correlation coefficient.

27. **Company Profits and Market Cap.** The scatter chart in the following figure was created using sample data for profits and market capitalizations from a sample of firms in the Fortune 500. **LO 13, 14**

DATA*file*
fortune500



    a. Discuss what the scatter chart indicates about the relationship between profits and market capitalization?
    b. The data used to produce this are contained in the file *fortune500*. Calculate the covariance between profits and market capitalization. Discuss what the covariance indicates about the relationship between profits and market capitalization?
    c. Calculate the correlation coefficient between profits and market capitalization. What does the correlation coefficient indicate about the relationship between profits and market capitalization?

28. **Jobless Rate and Percent of Delinquent Loans.** The economic downturn in 2008–2009 resulted in the loss of jobs and an increase in delinquent loans for housing. In projecting where the real estate market was headed in the coming year, economists studied the relationship between the jobless rate and the percentage of delinquent loans. The expectation was that if the jobless rate continued to increase, there would also be an increase in the percentage of delinquent loans. The following data show the jobless rate and the delinquent loan percentage for 27 major real estate markets. **LO 13, 14**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
joblessrate

Metro Area	Jobless Rate (%)	Delinquent Loans (%)	Metro Area	Jobless Rate (%)	Delinquent Loans (%)
Atlanta	7.1	7.02	New York	6.2	5.78
Boston	5.2	5.31	Orange County	6.3	6.08
Charlotte	7.8	5.38	Orlando	7.0	10.05
Chicago	7.8	5.40	Philadelphia	6.2	4.75
Dallas	5.8	5.00	Phoenix	5.5	7.22
Denver	5.8	4.07	Portland	6.5	3.79
Detroit	9.3	6.53	Raleigh	6.0	3.62
Houston	5.7	5.57	Sacramento	8.3	9.24
Jacksonville	7.3	6.99	St. Louis	7.5	4.40
Las Vegas	7.6	11.12	San Diego	7.1	6.91
Los Angeles	8.2	7.56	San Francisco	6.8	5.57
Miami	7.1	12.11	Seattle	5.5	3.87
Minneapolis	6.3	4.39	Tampa	7.5	8.42
Nashville	6.6	4.78			

Source: *The Wall Street Journal, January 27, 2009.*

a. Compute the correlation coefficient. Is there a positive correlation between the jobless rate and the percentage of delinquent housing loans? What is your interpretation?
b. Show a scatter chart of the relationship between the jobless rate and the percentage of delinquent housing loans.

Case Problem 1: Heavenly Chocolates Web Site Transactions

Heavenly Chocolates manufactures and sells quality chocolate products at its plant and retail store located in Saratoga Springs, New York. Two years ago, the company developed a web site and began selling its products over the Internet. Web site sales have exceeded the company's expectations, and management is now considering strategies to increase sales even further. To learn more about the web site customers, a sample of 50 Heavenly Chocolates transactions was selected from the previous month's sales. Data showing the day of the week each transaction was made, the type of browser the customer used, the time spent on the web site, the number of web pages viewed, and the amount spent by each of the 50 customers are contained in the file *heavenlychocolates*. A portion of the data is shown in the table that follows:

DATA*file*
heavenlychocolates

Customer	Day	Browser	Time (min)	Pages Viewed	Amount Spent ($)
1	Mon	Chrome	12.0	4	54.52
2	Wed	Other	19.5	6	94.90
3	Mon	Chrome	8.5	4	26.68
4	Tue	Firefox	11.4	2	44.73
5	Wed	Chrome	11.3	4	66.27
6	Sat	Firefox	10.5	6	67.80
7	Sun	Chrome	11.4	2	36.04
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
48	Fri	Chrome	9.7	5	103.15
49	Mon	Other	7.3	6	52.15
50	Fri	Chrome	13.4	3	98.75

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Heavenly Chocolates would like to use the sample data to determine whether online shoppers who spend more time and view more pages also spend more money during their visit to the web site. The company would also like to investigate the effect that the day of the week and the type of browser have on sales. **LO 2, 3, 5, 8, 9, 13, 14**

Managerial Report

Use the methods of descriptive statistics to learn about the customers who visit the Heavenly Chocolates web site. Include the following in your report.

1. Graphical and numerical summaries for the length of time the shopper spends on the web site, the number of pages viewed, and the mean amount spent per transaction. Discuss what you learn about Heavenly Chocolates' online shoppers from these numerical summaries.
2. Summarize the frequency, the total dollars spent, and the mean amount spent per transaction for each day of week. Discuss the observations you can make about Heavenly Chocolates' business based on the day of the week?
3. Summarize the frequency, the total dollars spent, and the mean amount spent per transaction for each type of browser. Discuss the observations you can make about Heavenly Chocolates' business based on the type of browser?
4. Develop a scatter chart, and compute the sample correlation coefficient to explore the relationship between the time spent on the web site and the dollar amount spent. Use the horizontal axis for the time spent on the web site. Discuss your findings.
5. Develop a scatter chart, and compute the sample correlation coefficient to explore the relationship between the number of web pages viewed and the amount spent. Use the horizontal axis for the number of web pages viewed. Discuss your findings.
6. Develop a scatter chart, and compute the sample correlation coefficient to explore the relationship between the time spent on the web site and the number of pages viewed. Use the horizontal axis to represent the number of pages viewed. Discuss your findings.

Case Problem 2: African Elephant Populations

Although millions of elephants once roamed across Africa, by the mid-1980s elephant populations in African nations had been devastated by poaching. Elephants are important to African ecosystems. In tropical forests, elephants create clearings in the canopy that encourage new tree growth. In savannas, elephants reduce bush cover to create an environment that is favorable to browsing and grazing animals. In addition, the seeds of many plant species depend on passing through an elephant's digestive tract before germination.

The status of the elephant now varies greatly across the continent. In some nations, strong measures have been taken to effectively protect elephant populations; for example, Kenya has destroyed over five tons of elephant ivory confiscated from poachers in an attempt to deter the growth of illegal ivory trade (Associated Press, July 20, 2011). In other nations the elephant populations remain in danger due to poaching for meat and ivory, loss of habitat, and conflict with humans. The table below shows elephant populations for several African nations in 1979, 1989, 2007, and 2012 (ElephantDatabase.org web site).

The David Sheldrick Wildlife Trust was established in 1977 to honor the memory of naturalist David Leslie William Sheldrick, who founded Warden of Tsavo East National Park in Kenya and headed the Planning Unit of the Wildlife Conservation and Management Department in that country. Management of the Sheldrick Trust would like to know what these data indicate about elephant populations in various African countries since 1979. **LO 2, 8, 12**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
africanelephants

Elephant Population

Country	1979	1989	2007	2012
Angola	12,400	12,400	2,530	2,530
Botswana	20,000	51,000	175,487	175,454
Cameroon	16,200	21,200	15,387	14,049
Cen African Rep	63,000	19,000	3,334	2,285
Chad	15,000	3,100	6,435	3,004
Congo	10,800	70,000	22,102	49,248
Dem Rep of Congo	377,700	85,000	23,714	13,674
Gabon	13,400	76,000	70,637	77,252
Kenya	65,000	19,000	31,636	36,260
Mozambique	54,800	18,600	26,088	26,513
Somalia	24,300	6,000	70	70
Tanzania	316,300	80,000	167,003	117,456
Zambia	150,000	41,000	29,231	21,589
Zimbabwe	30,000	43,000	99,107	100,291

Managerial Report

Use methods of descriptive statistics to summarize the data and comment on changes in elephant populations since 1979. Include the following in your report.

1. Use the geometric mean calculation to find the mean annual change in elephant population for each country in the 10 years from 1979 to 1989, and a discussion of which countries saw the largest changes in elephant population over this 10-year period.
2. Use the geometric mean calculation to find the mean annual change in elephant population for each country in the 18 years from 1989 to 2007, and a discussion of which countries saw the largest changes in elephant population over this 18-year period.
3. Use the geometric mean calculation to find the mean annual change in elephant population for each country in the five years from 2007 to 2012, and a discussion of which countries saw the largest changes in elephant population over this five-year period.
4. Create a multiple boxplot graph that includes boxplots of the elephant population observations in each year (1979, 1989, 2007, 2012). Use these boxplots and the results of your analyses in points 1 through 3 above to comment on how the populations of elephants have changed during these time periods.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 3

Data Visualization

Contents

Analytics in Action:
Cincinnati Zoo & Botanical Garden

3.1 Overview of Data Visualization
- Preattentive Attributes
- Data-Ink Ratio

3.2 Tables
- Table Design Principles
- Crosstabulation
- PivotTables in Excel

3.3 Charts
- Scatter Charts
- Recommended Charts in Excel
- Line Charts
- Bar Charts and Column Charts
- A Note on Pie Charts and Three-Dimensional Charts
- Additional Visualizations for Multiple Variables: Bubble Chart, Scatter Chart Matrix, and Table Lens
- PivotCharts in Excel

3.4 Specialized Data Visualizations
- Heat Maps
- Treemaps
- Waterfall Charts
- Stock Charts
- Parallel-Coordinates Chart

3.5 Visualizing Geospatial Data
- Choropleth Maps
- Cartograms

3.6 Data Dashboards
- Principles of Effective Data Dashboards
- Applications of Data Dashboards

Summary 134
Glossary 134
Problems 136

Available in the Cengage eBook:
- Appendix: Creating Tabular and Graphical Presentations with R
- Appendix: Data Visualization with Tableau

Learning Objectives

After completing this chapter, you will be able to

LO 1 Apply concepts of preattentive attributes and data-ink ratio to improve table and chart designs.

LO 2 Create and use PivotTables and PivotCharts to explore and analyze data.

LO 3 Create and interpret scatter charts to examine the relationship between two variables.

LO 4 Create and interpret line charts, bar charts, and column charts.

LO 5 Create and interpret trendlines fit to data on scatter charts and line charts.

LO 6 Create and interpret sorted bar (column) charts, clustered bar (column) charts, and stacked bar (column) charts.

LO 7 Create and interpret bubble charts, scatter chart matrixes, and table lenses to visualize the relationships among multiple variables.

LO 8 Create and interpret heat maps, sparklines, treemaps, waterfall charts, stock charts, and parallel coordinates plots to visualize data.

LO 9 Create and interpret choropleth maps and cartograms for applications to geospatial data.

LO 10 Apply the principles of effective data dashboards to suggest designs for data dashboards.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

Cincinnati Zoo & Botanical Garden[1]

The Cincinnati Zoo & Botanical Garden, located in Cincinnati, Ohio, is one of the oldest zoos in the United States. In 2019, it was named the best zoo in North America by *USA Today*. To improve decision making by becoming more data-driven, management decided they needed to link the various facets of their business and provide nontechnical managers and executives with an intuitive way to better understand their data. A complicating factor is that when the zoo is busy, managers are expected to be on the grounds interacting with guests, checking on operations, and dealing with issues as they arise or anticipating them. Therefore, being able to monitor what is happening in real time was a key factor in deciding what to do. Zoo management concluded that a data-visualization strategy was needed to address the problem.

Because of its ease of use, real-time updating capability, and iPad compatibility, the Cincinnati Zoo decided to implement its data-visualization strategy using IBM's Cognos advanced data-visualization software. Using this software, the Cincinnati Zoo developed the set of charts shown in Figure 3.1 (known as a data dashboard) to enable management to track the following key measures of performance:

- Item analysis (sales volumes and sales dollars by location within the zoo)
- Geoanalytics (using maps and displays of where the day's visitors are spending their time at the zoo)
- Customer spending

[1]The authors are indebted to John Lucas of the Cincinnati Zoo & Botanical Garden for providing this application.

Figure 3.1 Data Dashboard for the Cincinnati Zoo



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- Cashier sales performance
- Sales and attendance data versus weather patterns
- Performance of the zoo's loyalty rewards program

An iPad mobile application was also developed to enable the zoo's managers to be out on the grounds and still see and anticipate occurrences in real time. The Cincinnati Zoo's iPad application, shown in Figure 3.2, provides managers with access to the following information:

- Real-time attendance data, including what types of guests are coming to the zoo (members, nonmembers, school groups, etc.)
- Real-time analysis showing which locations are busiest and which items are selling the fastest inside the zoo
- Real-time geographical representation of where the zoo's visitors live

Having access to the data shown in Figures 3.1 and 3.2 allows the zoo managers to make better decisions about staffing levels, which items to stock based on weather and other conditions, and how to better target advertising based on geodemographics.

The impact that data visualization has had on the zoo has been substantial. Within the first year of use, the system was directly responsible for revenue growth of over $500,000, increased visitation to the zoo, enhanced customer service, and reduced marketing costs.

Figure 3.2 The Cincinnati Zoo iPad Data Dashboard



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The first step in trying to interpret data is often to visualize it in some way. Data visualization can be as simple as creating a summary table, or it could require generating charts to help interpret, analyze, and learn from the data. Data visualization is very helpful for identifying data errors and for reducing the size of your data set by highlighting important relationships and trends.

Data visualization is also important for conveying your analysis to others. Although business analytics is about making better decisions, in many cases, the ultimate decision maker is not the person who analyzes the data. Therefore, the person analyzing the data has to make the analysis simple for others to understand. Proper data-visualization techniques greatly improve the ability of the decision maker to interpret the analysis easily.

In this chapter, we discuss some general concepts related to data visualization including preattentive attributes and the data-ink ratio to help you analyze data and convey your analysis to others. We cover how to design tables and charts, as well as the most commonly used charts, and present an overview of some specialized data visualizations including those for use with geospatial data. We also discuss best practices for data dashboard design. Our detailed examples use Excel to generate tables and charts, and we discuss several software packages that can be used for advanced data visualization. The online appedixes explain how to use specialized software such as R and Tableau for data visualization.

3.1 Overview of Data Visualization

Decades of research studies in psychology and other fields show that the human mind can process visual images such as charts much faster than it can interpret rows and columns of numbers. However, these same studies also show that the human mind has certain limitations in its ability to interpret visual images and that some images are better at conveying information than others. The goal of this chapter is to introduce some of the most common forms of visualizing data and demonstrate when each form is appropriate.

Microsoft Excel is a ubiquitous tool used in business for basic data visualization. Software tools such as Excel make it easy for anyone to create many standard examples of data visualization. However, the default settings for tables and charts created with Excel can be altered to increase clarity. While there are a variety of software packages that are dedicated to data visualization, in this chapter we focus on the use of Excel. We provide online appendixes that introduce the use of R and Tableau for data visualization.

Preattentive Attributes

Effective data visualizations are easy and clear for the user to interpret. A well-designed data visualization reduces the **cognitive load**, or the amount of effort necessary to accurately and efficiently process the information being communicated by a data visualization. A data visualization with a reduced cognitive load is easier for the viewer to interpret.

Preattentive attributes are features that can be used in a data visualization to reduce the cognitive load required by the user to interpret the visualization. Preattentive attributes include attributes such as color, size, shape, and length among others. We can use a simple example to illustrate the power of preattentive attributes in data visualization. Examine Figure 3.3 and count the number of 7s in the figure as quickly as you can.

There are fourteen 7s in Figure 3.3. Even if you were able to find all fourteen 7s, it probably took you quite a bit of time, and it is likely that you made at least one mistake. Figures 3.4 and 3.5 demonstrate the use of preattentive attributes to make quickly counting the number of 7s much easier.

Figure 3.4 differentiates the 7s using color; each 7 is orange in color and all other numbers are black. This makes it much easier to identify the 7s, and it is much more likely that you can easily find all fourteen 7s quickly. Figure 3.5 uses size to differentiate the 7s. Because the 7s are larger than all the other numbers, it is much easier to find them quickly. Both color and size are preattentive attributes; these attributes reduce the cognitive load for

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.3 Count the Number of 7s that Appear in This Figure

7	3	4	1	3	4	5	6	4	0
3	0	6	9	0	4	5	8	6	3
2	7	2	2	9	9	4	5	2	1
2	2	4	5	2	0	9	2	0	4
2	4	0	7	6	9	3	0	0	4
7	7	8	9	2	6	7	2	4	7
6	1	3	3	2	1	4	4	9	0
3	6	6	2	7	5	5	2	5	4
1	1	4	0	6	3	4	0	5	1
3	7	5	2	7	5	7	7	3	9
3	3	8	6	9	5	5	3	6	4
7	6	0	3	0	9	9	0	2	9
4	6	9	4	8	2	6	5	8	3
9	3	9	2	2	8	4	3	9	8
5	8	8	2	9	1	2	4	8	5
1	7	4	0	1	1	9	9	5	8

Figure 3.4 Count the Number of 7s Aided by the Preattentive Attribute of Color

7	3	4	1	3	4	5	6	4	0
3	0	6	9	0	4	5	8	6	3
2	7	2	2	9	9	4	5	2	1
2	2	4	5	2	0	9	2	0	4
2	4	0	7	6	9	3	0	0	4
7	7	8	9	2	6	7	2	4	7
6	1	3	3	2	1	4	4	9	0
3	6	6	2	7	5	5	2	5	4
1	1	4	0	6	3	4	0	5	1
3	7	5	2	7	5	7	7	3	9
3	3	8	6	9	5	5	3	6	4
7	6	0	3	0	9	9	0	2	9
4	6	9	4	8	2	6	5	8	3
9	3	9	2	2	8	4	3	9	8
5	8	8	2	9	1	2	4	8	5
1	7	4	0	1	1	9	9	5	8

the users and makes it much easier to identify differences. This simple example shows the power of using preattentive attributes to convey meaning in data visualizations.

Figure 3.6 illustrates the three preattentive attributes discussed in this chapter: color, size, and length. In each case, the preattentive attribute reduces the required cognitive load and makes it much easier to distinguish differences in the visualization.

Color can be an extremely effective attribute to use to differentiate particular aspects of data in a visualization. However, one must be careful not to overuse color, as it can

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.5 Count the Number of 7s Aided by the Preattentive Attribute of Size

7 3 4 1 3 4 5 6 4 0
3 0 6 9 0 4 5 8 6 3

2 7 2 2 9 9 4 5 2 1
2 2 4 5 2 0 9 2 0 4

2 4 0 7 6 9 3 0 0 4
7 7 8 9 2 6 7 2 4 7
6 1 3 3 2 1 4 4 9 0

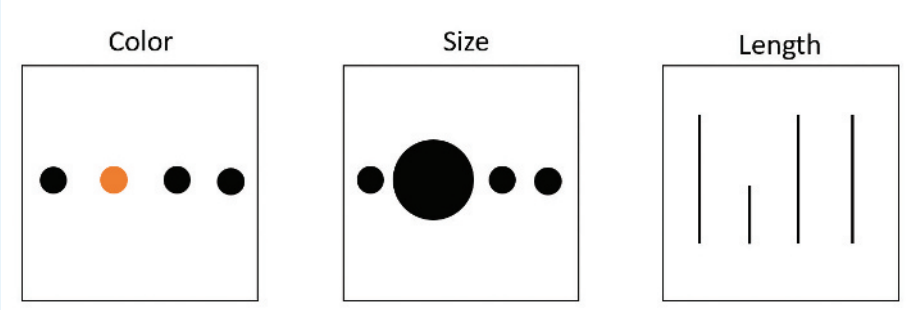
3 6 6 2 7 5 5 2 5 4
1 1 4 0 6 3 4 0 5 1

3 7 5 2 7 5 7 7 3 9
3 3 8 6 9 5 5 3 6 4

7 6 0 3 0 9 9 0 2 9
4 6 9 4 8 2 6 5 8 3
9 3 9 2 2 8 4 3 9 8
5 8 8 2 9 1 2 4 8 5

1 7 4 0 1 1 9 9 5 8

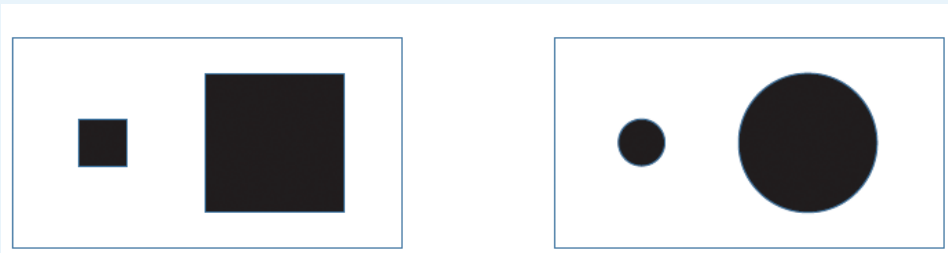
Figure 3.6 Preattentive Attributes Include Color, Size, and Shape



become distracting in a visualization. It should also be noted that many people suffer from color-blindness, which affects their ability to differentiate between some colors. Thus, we recommend a restrained use of color both to prevent distracting the user and also to avoid creating data visualizations that are not useful to those who are color-blind.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.7 Demonstration of Preattentive Attribute of Size



Size refers to the relative amount of 2D space that an object occupies in a visualization. One must be careful with the use of size in data visualizations because humans are not particularly good at judging relative differences in the 2D sizes of objects. Consider Figure 3.7, which shows a pair of squares and a pair of circles. Try to determine how much larger the bigger square and bigger circle are than their smaller counterparts.

In terms of area, both the bigger square and the bigger circle are nine times larger than their smaller counterparts. Most people are not good at estimating this relative size difference, so we must be careful when using the attribute of size to convey information about relative amounts. Because they rely on the use of size to convey relative amounts, the use of pie charts is often discouraged.

Length refers to the horizontal, vertical or diagonal distance of a line or bar/column. Length is useful for illustrating quantitative values because a longer line corresponds to a larger value. Length is used extensively in bar and column charts to visualize data. Because it is much easier to compare relative lengths than relative sizes, bar and column charts are often preferred to pie charts for visualizing data.

Data-Ink Ratio

The data-ink ratio was first described by Edward R. Tufte in his 2001 book The Visual Display of Quantitative Information.

Another helpful concept to keep in mind for creating effective tables and charts for data visualization is the idea of the **data-ink ratio**. The data-ink ratio measures the proportion of of ink used in a table or chart that is necessary to convey the meaning to the audience (known as "data-ink") to the total ink used for the table or chart. Non-data-ink is ink used in a table or chart that serves no useful purpose in conveying the data to the audience. Overuse of non-data-ink increases the cognitive load required by the user to interpret the chart or table. A guiding principle for effective data visualizations is that the table or graph should illustrate the data to help the user generate insights and understanding. The table or graph should not be so cluttered as to disguise the data or be difficult to interpret.

Let us consider the case of Gossamer Industries, a firm that produces fine silk clothing products. Gossamer is interested in tracking the sales of one of its most popular items, a particular style of scarf. Table 3.1 and Figure 3.8 provide examples of a table and chart with low data-ink ratios used to display sales of this style of scarf. The data used in this table and figure represent product sales by day. Both of these examples are similar to tables and charts generated with Excel using common default settings. In Table 3.1, most of the grid lines serve no useful purpose. Likewise, in Figure 3.8, the horizontal lines in the chart also add little additional information. In both cases, most of these lines can be deleted without reducing the information conveyed. However, an important piece of information is missing from Figure 3.8: labels for axes. Unless both the meaning and the unit of measure are obvious, axes should always be labeled.

Table 3.2 shows a modified table in which all grid lines have been deleted except for those around the title of the table. Deleting the grid lines in Table 3.1 increases the data-ink ratio because a larger proportion of the ink in the table is used to convey the information

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

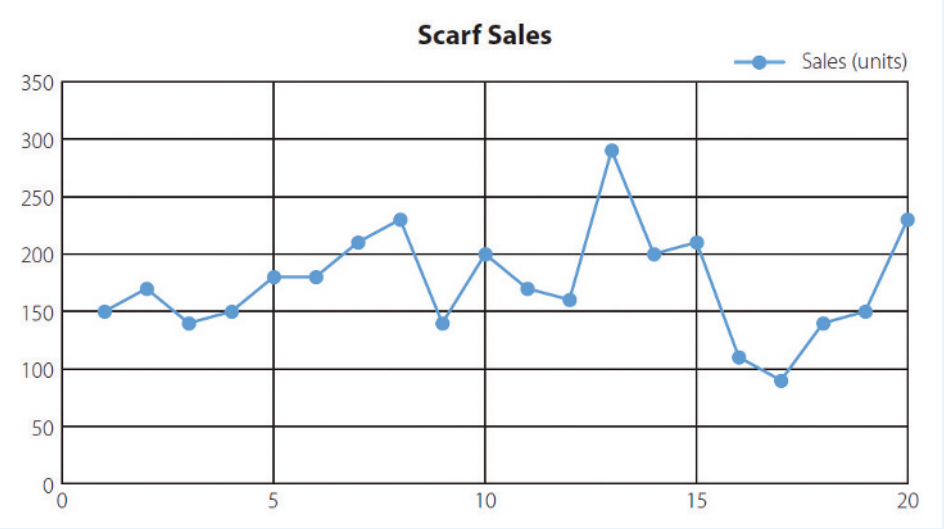
Table 3.1 Example of a Low Data-Ink Ratio Table

Scarf Sales			
Day	**Sales (units)**	**Day**	**Sales (units)**
1	150	11	170
2	170	12	160
3	140	13	290
4	150	14	200
5	180	15	210
6	180	16	110
7	210	17	90
8	230	18	140
9	140	19	150
10	200	20	230

(the actual numbers). Similarly, deleting the unnecessary horizontal and vertical grid lines in Figure 3.8, as shown in Figure 3.9, increases the data-ink ratio. Note that deleting these horizontal grid lines and removing (or reducing the size of) the markers at each data point can make it more difficult to determine the exact values plotted in the chart. However, as we discuss later, a simple chart is not the most effective way of presenting data when the audience needs to know exact values; in these cases, it is better to use a table.

In many cases, white space in a table or a chart can improve readability. This principle is similar to the idea of increasing the data-ink ratio. Consider Table 3.2 and Figure 3.9. Removing the unnecessary lines has increased the "white space," making it easier to read both the table and the chart. The process of increasing the data-ink ratio by adding whitespace is also known as **decluttering**. The fundamental idea in creating effective tables and charts is to make them as simple as possible in conveying information to the reader.

Figure 3.8 Example of a Low Data-Ink Ratio Chart

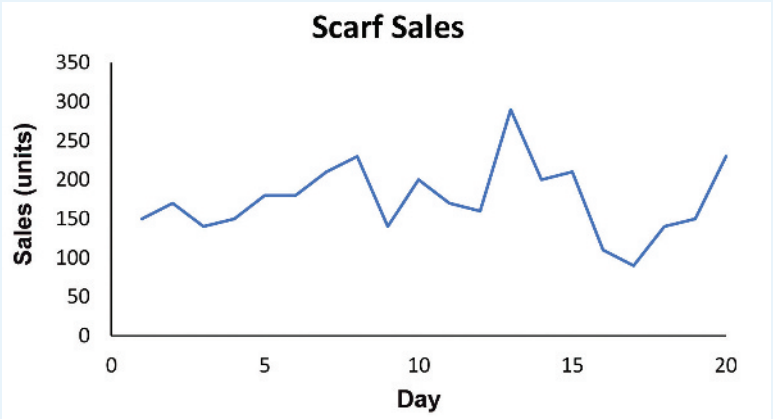


Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 3.2 Increasing the Data-Ink Ratio by Removing Unnecessary Gridlines

Scarf Sales			
Day	**Sales (units)**	**Day**	**Sales (units)**
1	150	11	170
2	170	12	160
3	140	13	290
4	150	14	200
5	180	15	210
6	180	16	110
7	210	17	90
8	230	18	140
9	140	19	150
10	200	20	230

Figure 3.9 Increasing the Data-Ink Ratio by Adding Labels to Axes and Removing Unnecessary Lines and Labels



Notes + Comments

1. Tables have been used to display data for more than a thousand years. However, charts are much more recent inventions. The famous 17th-century French mathematician, René Descartes, is credited with inventing the now familiar graph with horizontal and vertical axes. William Playfair invented bar charts, line charts, and pie charts in the late 18th century, all of which we will discuss in this chapter. More recently, individuals such as William Cleveland, Edward R. Tufte, and Stephen Few have introduced design techniques for both clarity and beauty in data visualization.
2. Many of the default settings in Excel are not ideal for displaying data using tables and charts. Before presenting Excel-generated tables and charts to others, it is worth the effort to remove unnecessary lines and labels.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3.2 Tables

The first decision in displaying data is whether a table or a chart will be more effective. In general, charts can often convey information faster and easier to readers, but in some cases a table is more appropriate. Tables should be used when

1. the reader needs to refer to specific numerical values.
2. the reader needs to make precise comparisons between different values and not just relative comparisons.
3. the values being displayed have different units or very different magnitudes.

When the accounting department of Gossamer Industries is summarizing the company's annual data for completion of its federal tax forms, the specific numbers corresponding to revenues and expenses are important rather than just the relative values. Therefore, these data should be presented in a table similar to Table 3.3.

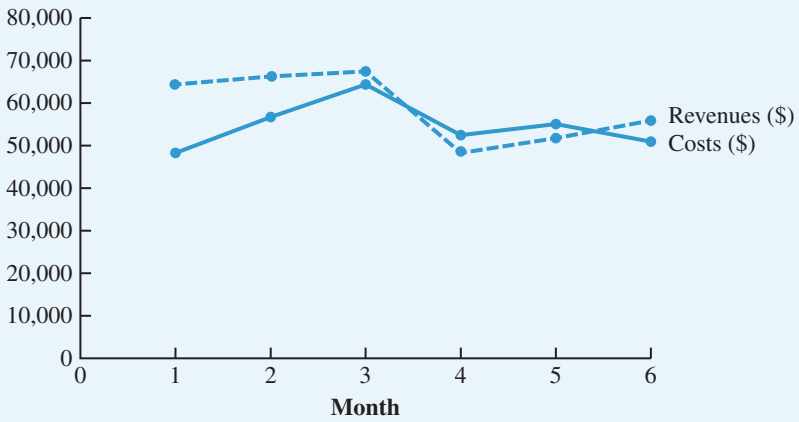
Similarly, if it is important to know by exactly how much revenues exceed expenses each month, then this would also be better presented as a table rather than as a line chart as seen in Figure 3.10. Notice that it is very difficult to determine the monthly revenues and costs in Figure 3.10. We could add these values using data labels, but they would clutter the figure. The preferred solution is to combine the chart with the table into a single figure, as in Figure 3.11. This allows the reader to easily see the monthly changes in revenues and costs while also being able to refer to the exact numerical values.

Now suppose that you wish to display data on revenues, costs, and head count for each month. Costs and revenues are measured in dollars, but head count is measured in number of employees. Although all these values can be displayed on a line chart using multiple

Table 3.3 Table Showing Exact Values for Costs and Revenues by Month for Gossamer Industries

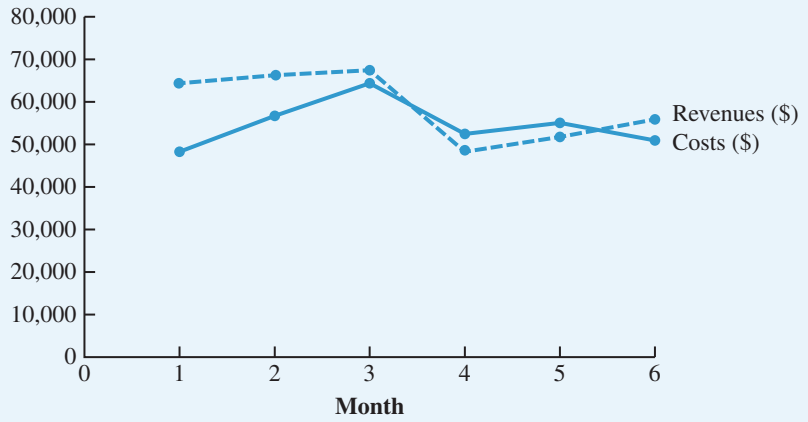
	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027

Figure 3.10 Line Chart of Monthly Costs and Revenues at Gossamer Industries



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.11 Combined Line Chart and Table for Monthly Costs and Revenues at Gossamer Industries



	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027

vertical axes, this is generally not recommended. Because the values have widely different magnitudes (costs and revenues are in the tens of thousands, whereas head count is approximately 10 each month), it would be difficult to interpret changes on a single chart. Therefore, a table similar to Table 3.4 is recommended.

Table Design Principles

In designing an effective table, keep in mind the data-ink ratio and avoid the use of unnecessary ink in tables. In general, this means that we should avoid using vertical lines in a table unless they are necessary for clarity. Horizontal lines are generally necessary only for separating column titles from data values or when indicating that a calculation has taken place. Consider Figure 3.12, which compares several forms of a table displaying Gossamer's costs and revenue data. Most people find Design D, with the fewest grid lines, easiest to read. In this table, grid lines are used only to separate the column headings from the data and to indicate that a calculation has occurred to generate the Profits row and the Total column.

In large tables, vertical lines or light shading can be useful to help the reader differentiate the columns and rows. Table 3.5 breaks out the revenue data by location for nine cities

Table 3.4 Table Displaying Head Count, Costs, and Revenues at Gossamer Industries

	Month						
	1	2	3	4	5	6	Total
Head count	8	9	10	9	9	9	
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.12 Comparing Different Table Designs

Design A:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design B:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design C:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design D:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

and shows 12 months of revenue and cost data. In Table 3.5, every other column has been lightly shaded. This helps the reader quickly scan the table to see which values correspond with each month. The horizontal line between the revenue for Academy and the Total row helps the reader differentiate the revenue data for each location and indicates that a calculation has taken place to generate the totals by month. If one wanted to highlight the differences among locations, the shading could be done for every other row instead of every other column.

We depart from these guidelines in some figures and tables in this textbook to more closely match Excel's output.

Notice also the alignment of the text and numbers in Table 3.5. Columns of numerical values in a table should usually be right-aligned; that is, the final digit of each number should be aligned in the column. This makes it easy to see differences in the magnitude of values. If you are showing digits to the right of the decimal point, all values should include the same number of digits to the right of the decimal. Also, use only the number of digits that are necessary to convey the meaning in comparing the values; there is no need to include additional digits if they are not meaningful for comparisons. In many business applications, we report financial values, in which case we often round to the nearest dollar or include two digits to the right of the decimal if such precision is necessary. Additional digits to the right of the decimal are usually unnecessary. For extremely large numbers, we may prefer to display data rounded to the nearest thousand, ten thousand, or even million. For instance, if we need to include, say, $3,457,982 and $10,124,390 in a table when exact dollar values are not necessary, we could write these as 3.458 and 10.124 and indicate that all values in the table are in units of $1,000,000, usually by including "($ millions)" in the title of the column.

It is generally best to left-align text values within a column in a table, as in the Revenues by Location (the first) column of Table 3.5. In some cases, you may prefer to center text, but you should do this only if the text values are all approximately the same length. Otherwise, aligning the first letter of each data entry promotes readability. Column headings should either match the alignment of the data in the columns or be centered over the values, as in Table 3.5.

Crosstabulation

A useful approach for describing data of two variables is a **crosstabulation**, which is a tabular summary of data for two variables. To illustrate, consider the following application based on data from Zagat's Restaurant Review. Data on the quality rating, meal price, and the usual wait time for a table during peak hours were collected for a sample of 300 Los Angeles area restaurants. Table 3.6 shows the data for the first 10 restaurants.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 3.5 Larger Table Showing Revenues by Location for 12 Months of Data

	Month												
Revenues by Location ($)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
Temple	8,987	8,595	8,958	6,718	8,066	8,574	8,701	9,490	9,610	9,262	9,875	11,058	107,895
Killeen	8,212	9,143	8,714	6,869	8,150	8,891	8,766	9,193	9,603	10,374	10,456	10,982	109,353
Waco	11,603	12,063	11,173	9,622	8,912	9,553	11,943	12,947	12,925	14,050	14,300	13,877	142,967
Belton	7,671	7,617	7,896	6,899	7,877	6,621	7,765	7,720	7,824	7,938	7,943	7,047	90,819
Granger	7,642	7,744	7,836	5,833	6,002	6,728	7,848	7,717	7,646	7,620	7,728	8,013	88,357
Harker Heights	5,257	5,326	4,998	4,304	4,106	4,980	5,084	5,061	5,186	5,179	4,955	5,326	59,763
Gatesville	5,316	5,245	5,056	3,317	3,852	4,026	5,135	5,132	5,052	5,271	5,304	5,154	57,859
Lampasas	5,266	5,129	5,022	3,022	3,088	4,289	5,110	5,073	4,978	5,343	4,984	5,315	56,620
Academy	4,170	5,266	7,472	1,594	1,732	2,025	8,772	1,956	3,304	3,090	3,579	2,487	45,446
Total	64,124	66,128	67,125	48,178	51,785	55,687	69,125	64,288	66,128	68,128	69,125	69,258	759,079
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	57,898	62,050	65,215	61,819	67,828	69,558	710,935

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 3.6 Quality Rating and Meal Price for 300 Los Angeles Restaurants

DATA*file* restaurant

Restaurant	Quality Rating	Meal Price ($)	Wait Time (min)
1	Good	18	5
2	Very Good	22	6
3	Good	28	1
4	Excellent	38	74
5	Very Good	33	6
6	Good	28	5
7	Very Good	19	11
8	Very Good	11	9
9	Very Good	23	13
10	Good	13	1

Types of data such as categorical and quantitative are discussed in Chapter 2.

Quality ratings are an example of categorical data, and meal prices are an example of quantitative data.

For now, we will limit our consideration to the quality-rating and meal-price variables. A crosstabulation of the data for quality rating and meal price is shown in Table 3.7. The left and top margin labels define the classes for the two variables. In the left margin, the row labels (Good, Very Good, and Excellent) correspond to the three classes of the quality-rating variable. In the top margin, the column labels ($10–19, $20–29, $30–39, and $40–49) correspond to the four classes (or bins) of the meal-price variable. Each restaurant in the sample provides a quality rating and a meal price. Thus, each restaurant in the sample is associated with a cell appearing in one of the rows and one of the columns of the crosstabulation. For example, restaurant 5 is identified as having a very good quality rating and a meal price of $33. This restaurant belongs to the cell in row 2 and column 3. In constructing a crosstabulation, we simply count the number of restaurants that belong to each of the cells in the crosstabulation.

Table 3.7 shows that the greatest number of restaurants in the sample (64) have a very good rating and a meal price in the $20–29 range. Only two restaurants have an excellent rating and a meal price in the $10–19 range. Similar interpretations of the other frequencies can be made. In addition, note that the right and bottom margins of the crosstabulation give the frequencies of quality rating and meal price separately. From the right margin, we see that data on quality ratings show 84 good restaurants, 150 very good restaurants, and 66 excellent restaurants. Similarly, the bottom margin shows the counts for the meal price variable. The value of 300 in the bottom-right corner of the table indicates that 300 restaurants were included in this data set.

Table 3.7 Crosstabulation of Quality Rating and Meal Price for 300 Los Angeles Restaurants

	Meal Price				
Quality Rating	$10–19	$20–29	$30–39	$40–49	Total
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

PivotTables in Excel

A crosstabulation in Microsoft Excel is known as a **PivotTable**. We will first look at a simple example of how Excel's PivotTable is used to create a crosstabulation of the Zagat's restaurant data shown previously. Figure 3.13 illustrates a portion of the data contained in the file *restaurant*; the data for the 300 restaurants in the sample have been entered into cells B2:D301.

To create a PivotTable in Excel, we follow these steps:

DATA *file*
restaurant

Step 1. Click the **Insert** tab on the Ribbon

Step 2. Click the **PivotTable** button in the **Tables** group

Step 3. When the **Create PivotTable** dialog box appears:
Choose **Select a table or range**
Enter *A1:D301* in the **Table/Range:** box
Select **New Worksheet** as the location for the PivotTable Report
Click **OK**

The resulting initial PivotTable Field List and PivotTable Report are shown in Figure 3.14.

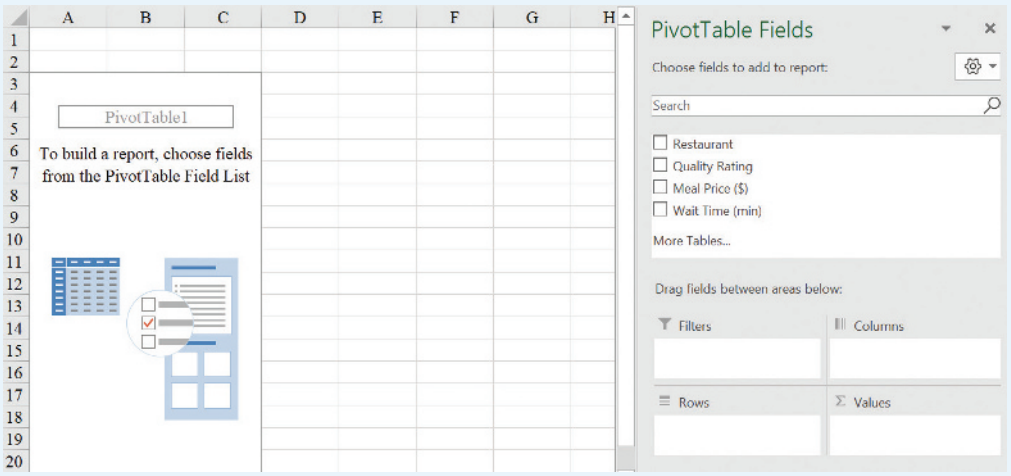
Each of the four columns in Figure 3.13 [Restaurant, Quality Rating, Meal Price ($), and Wait Time (min)] is considered a field by Excel. Fields may be chosen to represent rows, columns, or values in the body of the PivotTable Report. The following steps show how to use Excel's PivotTable Field List to assign the Quality Rating field to the rows, the Meal Price ($) field to the columns, and the Restaurant field to the body of the PivotTable report.

Figure 3.13 Excel Worksheet Containing Restaurant Data

	A	B	C	D
1	Restaurant	Quality Rating	Meal Price ($)	Wait Time (min)
2	1	Good	18	5
3	2	Very Good	22	6
4	3	Good	28	1
5	4	Excellent	38	74
6	5	Very Good	33	6
7	6	Good	28	5
8	7	Very Good	19	11
9	8	Very Good	11	9
10	9	Very Good	23	13
11	10	Good	13	1
12	11	Very Good	33	18
13	12	Very Good	44	7
14	13	Excellent	42	18
15	14	Excellent	34	46
16	15	Good	25	0
17	16	Good	22	3
18	17	Good	26	3
19	18	Excellent	17	36
20	19	Very Good	30	7
21	20	Good	19	3
22	21	Very Good	33	10
23	22	Very Good	22	14
24	23	Excellent	32	27
25	24	Excellent	33	80
26	25	Very Good	34	9

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.14 Initial PivotTable Field List and PivotTable Field Report for the Restaurant Data



Step 4. In the **PivotTable Fields** task pane, go to **Drag fields between areas below:**
Drag the **Quality Rating** field to the **Rows** area
Drag the **Meal Price ($)** field to the **Columns** area
Drag the **Restaurant** field to the **Values** area
Step 5. Click on **Sum of Restaurant** in the **Values** area
Step 6. Select **Value Field Settings** from the list of options
Step 7. When the **Value Field Settings** dialog box appears:
Under **Summarize value field by**, select **Count**
Click **OK**

Figure 3.15 shows the completed PivotTable Field List and a portion of the PivotTable worksheet as it now appears.

To complete the PivotTable, we need to group the columns representing meal prices and place the row labels for quality rating in the proper order:

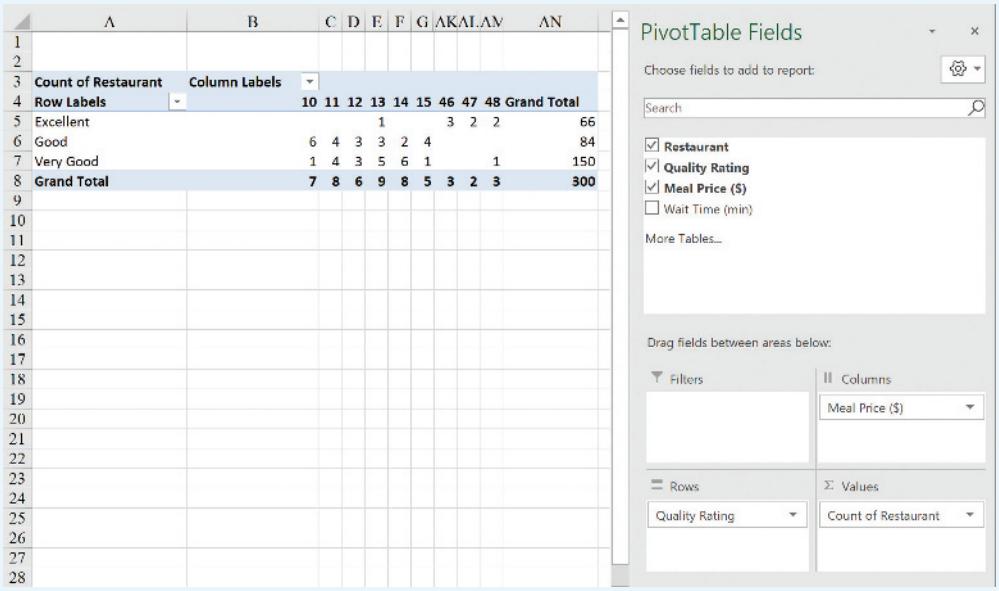
Step 8. Right-click in cell B4 or any other cell containing a meal price column label
Step 9. Select **Group** from the list of options
Step 10. When the **Grouping** dialog box appears:
Enter *10* in the **Starting at:** box
Enter *49* in the **Ending at:** box
Enter *10* in the **By:** box
Click **OK**
Step 11. Right-click on "Excellent" in cell A5
Step 12. Select **Move** and click **Move "Excellent" to End**

The final PivotTable, shown in Figure 3.16, provides the same information as the crosstabulation in Table 3.7.

The values in Figure 3.16 can be interpreted as the frequencies of the data. For instance, row 8 provides the frequency distribution for the data over the quantitative variable of meal price. Seventy-eight restaurants have meal prices of $10 to $19. Column F provides the frequency distribution for the data over the categorical variable of quality.

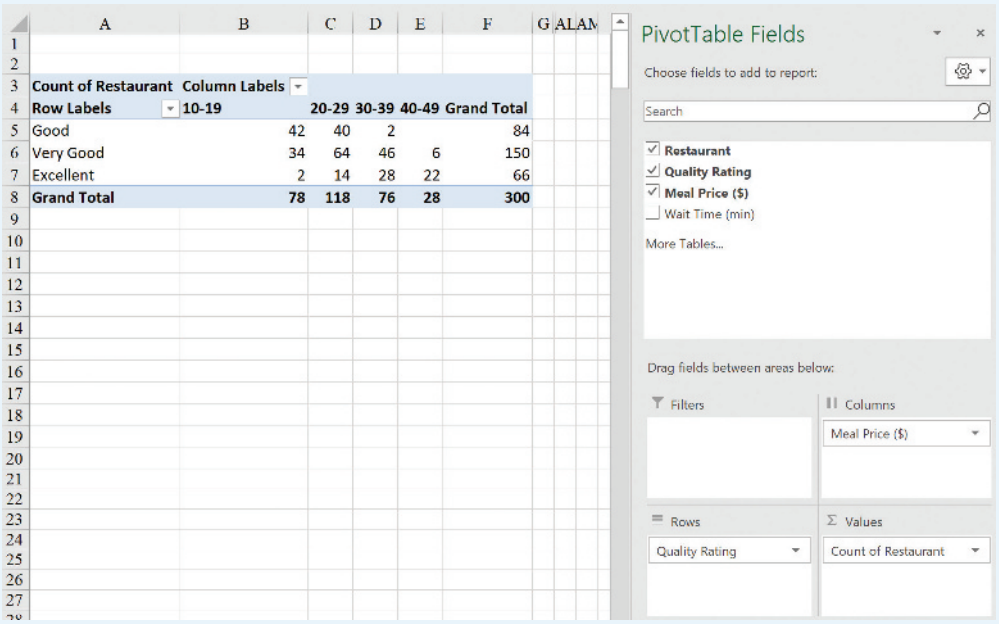
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.15 Completed PivotTable Field List and a Portion of the PivotTable Report for the Restaurant Data (Columns H:AK Are Hidden)



Count of Restaurant	Column Labels									
Row Labels	10	11	12	13	14	15	46	47	48	Grand Total
Excellent				1			3	2	2	66
Good	6	4	3	3	2	4				84
Very Good	1	4	3	5	6	1			1	150
Grand Total	7	8	6	9	8	5	3	2	3	300

Figure 3.16 Final PivotTable Report for the Restaurant Data



Count of Restaurant	Column Labels				
Row Labels	10-19	20-29	30-39	40-49	Grand Total
Good	42	40	2		84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Grand Total	78	118	76	28	300

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A total of 150 restaurants have a quality rating of Very Good. We can also use a PivotTable to create percent frequency distributions, as shown in the following steps:

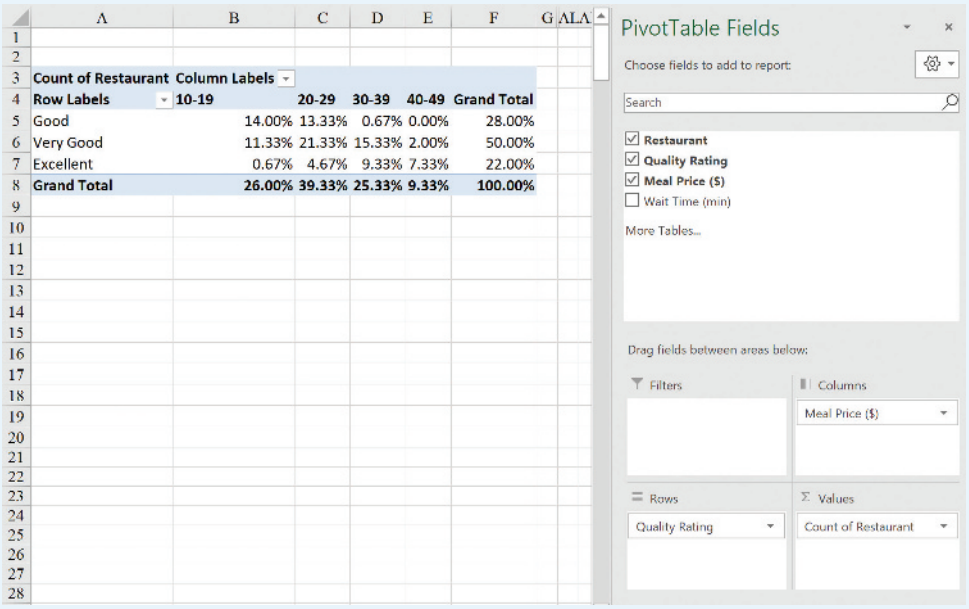
Step 1. To invoke the **PivotTable Fields** task pane, select any cell in the pivot table
Step 2. In the **PivotTable Fields** task pane, click the **Count of Restaurant** in the **Values** area
Step 3. Select **Value Field Settings . . .** from the list of options
Step 4. When the **Value Field Settings** dialog box appears, click the tab for **Show Values As**
Step 5. In the **Show values as** area, select **% of Grand Total** from the drop-down menu
Click **OK**

Figure 3.17 displays the percent frequency distribution for the Restaurant data as a PivotTable. The figure indicates that 50% of the restaurants are in the Very Good quality category and that 26% have meal prices between $10 and $19.

PivotTables in Excel are interactive, and they may be used to display statistics other than a simple count of items. As an illustration, we can easily modify the PivotTable in Figure 3.16 to display summary information on wait times instead of meal prices.

Step 1. To invoke the **PivotTable Fields** task pane, select any cell in the pivot table
Step 2. In the **PivotTable Fields** task pane, click the **Count of Restaurant** field in the **Values** area
Select **Remove Field**
Step 3. Drag the **Wait Time (min)** to the **Values** area
Step 4. Click on **Sum of Wait Time (min)** in the **Values** area
Step 5. Select **Value Field Settings…** from the list of options
Step 6. When the **Value Field Settings** dialog box appears:
Under **Summarize value field by**, select **Average**
Click **Number Format**
In the **Category:** area, select **Number**
Enter *1* for **Decimal places:**
Click **OK**
When the **Value Field Settings** dialog box reappears, click **OK**

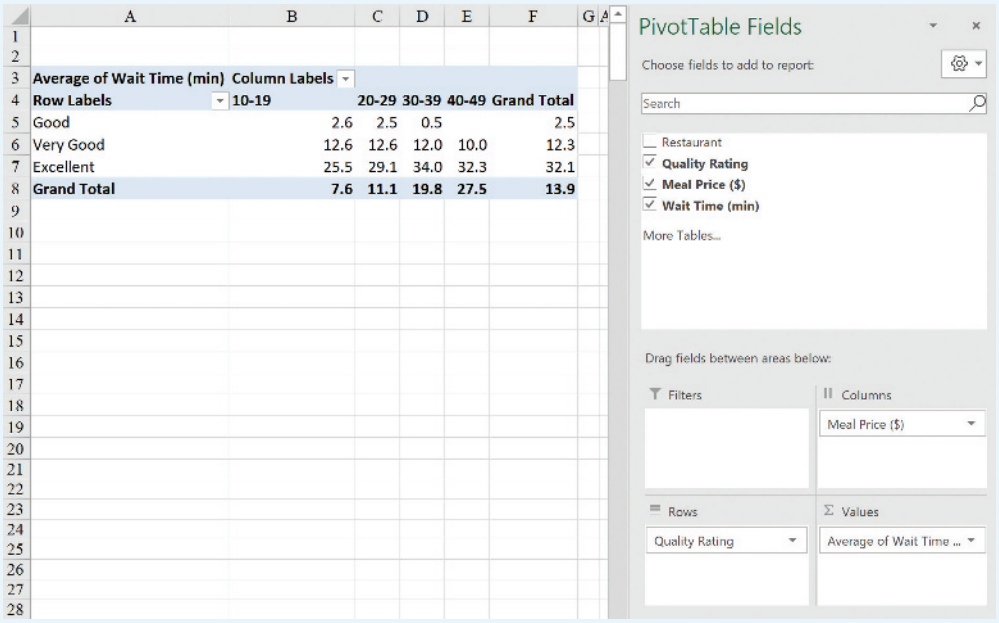
Figure 3.17 Percent Frequency Distribution as a PivotTable for the Restaurant Data



Count of Restaurant	Column Labels				
Row Labels	10-19	20-29	30-39	40-49	Grand Total
Good	14.00%	13.33%	0.67%	0.00%	28.00%
Very Good	11.33%	21.33%	15.33%	2.00%	50.00%
Excellent	0.67%	4.67%	9.33%	7.33%	22.00%
Grand Total	26.00%	39.33%	25.33%	9.33%	100.00%

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.18 PivotTable Report for the Restaurant Data with Average Wait Times Added



Average of Wait Time (min)	Column Labels				
Row Labels	10-19	20-29	30-39	40-49	Grand Total
Good	2.6	2.5	0.5		2.5
Very Good	12.6	12.6	12.0	10.0	12.3
Excellent	25.5	29.1	34.0	32.3	32.1
Grand Total	**7.6**	**11.1**	**19.8**	**27.5**	**13.9**

The completed PivotTable appears in Figure 3.18. This PivotTable replaces the counts of restaurants with values for the average wait time for a table at a restaurant for each grouping of meal prices ($10–19, $20–29, $30–39, and $40–49). For instance, cell B7 indicates that the average wait time for a table at an Excellent restaurant with a meal price of $10–19 is 25.5 minutes. Column F displays the total average wait times for tables in each quality rating category. We see that Excellent restaurants have the longest average wait of 35.2 minutes and that Good restaurants have an average wait time of only 2.5 minutes. Finally, cell D7 shows us that the longest wait times can be expected at Excellent restaurants with meal prices in the $30–39 range (34 minutes).

*You can also filter data in a PivotTable by dragging the field that you want to filter to the **Filters** area in the **PivotTable Fields**.*

We can also examine only a portion of the data in a PivotTable using the Filter option in Excel. To Filter data in a PivotTable, click on the **Filter Arrow** next to **Row Labels** or **Column Labels** and then uncheck the values that you want to remove from the PivotTable. For example, we could click on the arrow next to Row Labels and then uncheck the Good value to examine only Very Good and Excellent restaurants.

3.3 Charts

Charts (or graphs) are visual methods for displaying data. In this section, we introduce some of the most commonly used charts to display and analyze data, including scatter charts, line charts, and bar charts. Excel is a popular software package for creating simple charts. We explain how to use Excel to create scatter charts, line charts, sparklines, bar charts, and column charts.

Scatter Charts

A **scatter chart** is a graphical presentation of the relationship between two quantitative variables. As an illustration, consider the advertising/sales relationship for an electronics store in San Francisco. On 10 occasions during the past three months, the store used

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 3.8 Sample Data for the San Francisco Electronics Store

	x	*y*
Week	Number of Commercials	Sales ($100s)
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

DATA*file* electronics

weekend television commercials to promote sales at its stores. The managers want to investigate whether a relationship exists between the number of commercials shown and sales at the store the following week. Sample data for the 10 weeks, with sales in hundreds of dollars, are shown in Table 3.8.

We will use the data from Table 3.8 to create a scatter chart using Excel's chart tools and the data in the file *electronics*:

Step 1. Select cells B1:C11

Step 2. Click the **Insert** tab on the Ribbon
In the **Charts** group, click **Insert Scatter (X,Y) or Bubble Chart**

Step 3. When the list of scatter diagram subtypes appears:
Click **Scatter**

Step 4. Click the **Chart Title** and replace it with *Relationship Between Sales and Number of Commercials for San Francisco Electronics Store*

Step 5. Click the **Chart Elements** button (located next to the top right corner of the chart)

Step 6. When the list of chart elements appears:
Select the check box for **Axis Titles** (to create placeholders for the axis titles)
Deselect the check box for **Gridlines** (to remove the grid lines from the chart)
Select the check box for **Trendline** (to add a trendline to the chart)

Steps 6 through 8 improve the data-ink ratio by adding axes titles and removing unnecessary gridlines.

Step 7. Click the horizontal **Axis Title** placeholder and replace it with *Number of Commercials*

Step 8. Click the vertical **Axis Title** placeholder and replace it with *Sales ($100s)*

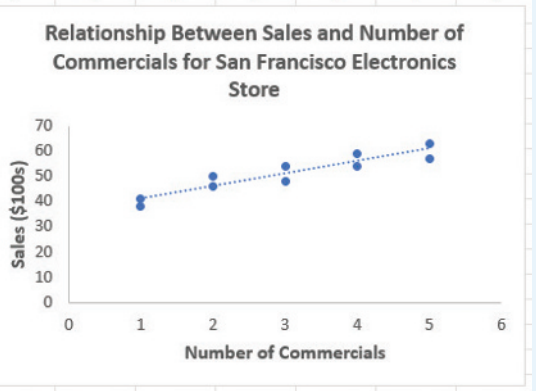
In step 6 we use Excel to add a trendline to the scatter chart. A **trendline** is a line that provides an approximation of the relationship between the variables.

Figure 3.19 shows the scatter chart and linear trendline created with Excel for the data in Table 3.8. The number of commercials (x) is shown on the horizontal axis, and sales (y) are shown on the vertical axis. For week 1, $x = 2$ and $y = 50$. A point is plotted on the scatter chart at those coordinates; similar points are plotted for the other nine weeks. Note that during two of the weeks, one commercial was shown, during two of the weeks, two commercials were shown, and so on.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.19 Scatter Chart for the San Francisco Electronics Store

	A	B	C
1	Week	Number of Commercials	Sales ($100s)
2	1	2	50
3	2	5	57
4	3	1	41
5	4	3	54
6	5	4	54
7	6	1	38
8	7	5	63
9	8	3	48
10	9	4	59
11	10	2	46



Scatter charts are also referred to as scatter plots or scatter diagrams.

Chapter 2 introduces scatter charts and relates them to the concepts of covariance and correlation.

The completed scatter chart in Figure 3.19 indicates a positive linear relationship (or positive correlation) between the number of commercials and sales: higher sales are associated with a higher number of commercials. The linear relationship is not perfect because not all of the points are on a straight line. However, the general pattern of the points and the trendline suggest that the overall relationship is positive. This implies that the covariance between sales and commercials is positive and that the correlation coefficient between these two variables is between 0 and +1.

The **Chart Buttons** in Excel allow users to quickly modify and format charts. Three buttons appear next to a chart whenever you click on a chart to make it active. Clicking on the **Chart Elements** button brings up a list of check boxes to quickly add and remove axes, axis titles, chart titles, data labels, trendlines, and more. Clicking on the **Chart Styles** button allows the user to quickly choose from many preformatted styles to change the look of the chart. Clicking on the **Chart Filter** button allows the user to select the data to be included in the chart. The Chart Filter button is very useful for performing additional data analysis.

Recommended Charts in Excel

Excel has the ability to recommend charts for a given data set. The steps below demonstrate the Recommended Charts tool in Excel for the San Francisco electronics store data.

DATA*file* electronics

Step 1. Select cells B1:C11
Step 2: Click the **Insert** tab in the Ribbon
Step 3: Click the **Recommended Charts** button in the **Charts** group
Step 4: When the **Insert Chart** dialog box appears, select the **Scatter** option (see Figure 3.20)
Click **OK**

These steps create the basic scatter chart that can then be formatted (using the Chart Buttons) to create the completed scatter chart shown in Figure 3.19. Note that the Recommended Charts tool gives several possible recommendations for the electronics

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

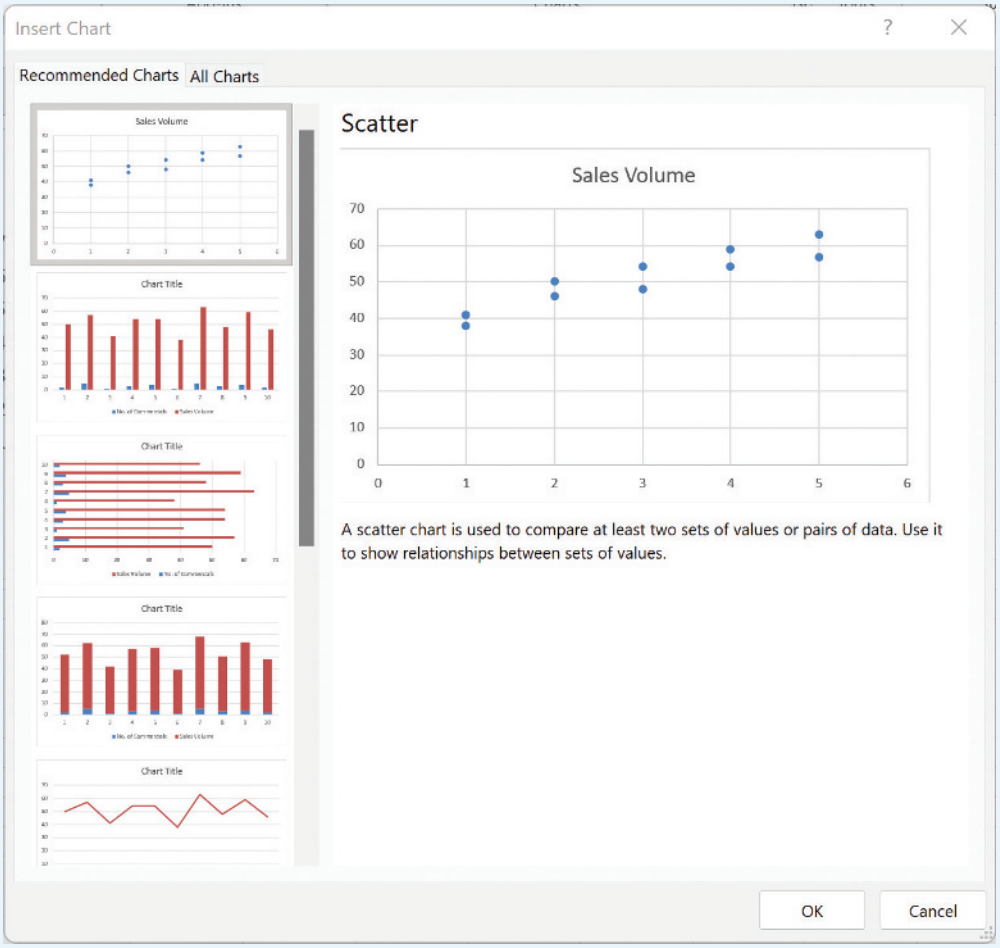
data in Figure 3.20. In addition to a scatter chart, these recommendations include a line chart and a bar chart, and bar charts, which are covered later in this chapter. Excel's Recommended Charts tool generally does a good job of interpreting your data and providing recommended charts, but take care to ensure that the selected chart is meaningful and follows good design practice.

Line Charts

A line chart for time series data is often called a time series plot.

Line charts are similar to scatter charts, but a line connects the points in the chart. Line charts are very useful for time series data collected over a period of time (minutes, hours, days, years, etc.). As an example, Kirkland Industries sells air compressors to manufacturing companies. Table 3.9 contains total sales amounts (in $100s) for air compressors during each month in the most recent calendar year. Figure 3.21 displays a scatter chart and a line chart created in Excel for these sales data. The line chart connects the points of the scatter chart. The addition of lines between the points suggests continuity, and it is easier for the reader to interpret changes over time.

Figure 3.20 Insert Chart Dialog Box from Recommended Charts Tool in Excel



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
kirkland

Table 3.9 Monthly Sales Data of Air Compressors at Kirkland Industries

Month	Sales ($100s)
Jan	135
Feb	145
Mar	175
Apr	180
May	160
Jun	135
Jul	210
Aug	175
Sep	160
Oct	120
Nov	115
Dec	120

To create the line chart in Figure 3.21 in Excel, we follow these steps:

Step 1. Select cells A1:B13

Step 2. Click the **Insert** tab on the Ribbon

Step 3. Click the **Insert Line Chart** button in the **Charts** group

Step 4. When the list of line chart subtypes appears, click the **Line** button under **2-D Line**

This creates a line chart for sales with a basic layout and minimum formatting

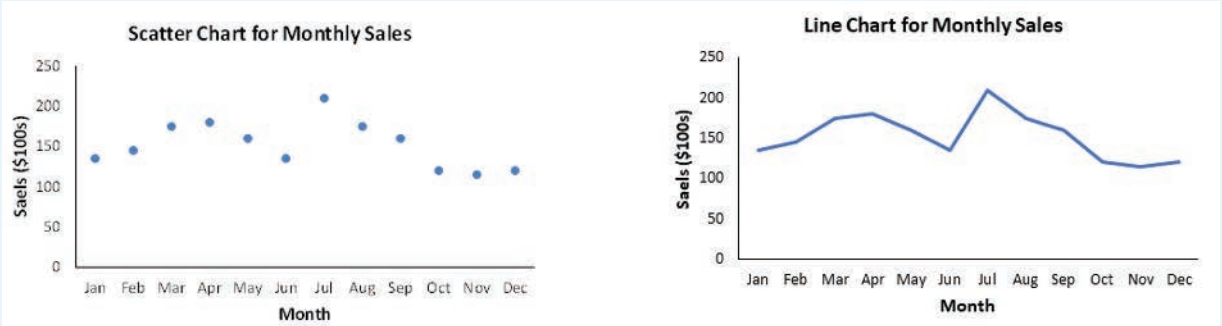
Step 5. Select the line chart that was just created to reveal the **Chart Buttons**

Step 6. Click the **Chart Elements** button

Select the check boxes for **Axes**, **Axis Titles**, and **Chart Title**

Deselect the check box for **Gridlines**

Figure 3.21 Scatter Chart and Line Chart for Monthly Sales Data at Kirkland Industries



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

*Because the gridlines do not add any meaningful information here, we do not select the check box for **Gridlines** in **Chart Elements**, because omitting the gridlines increases the data-ink ratio.*

Click on the text box next to the vertical axis, and replace "Axis Title" with *Sales ($100s)*
Click on the text box next to the horizontal axis and replace "Axis Title" with *Month*
Click on the text box above the chart, and replace "Sales ($100s)" with *Monthly Sales of Kirkland Air Compressors*

Figure 3.22 shows the line chart created in Excel along with the selected options for the Chart Elements button.

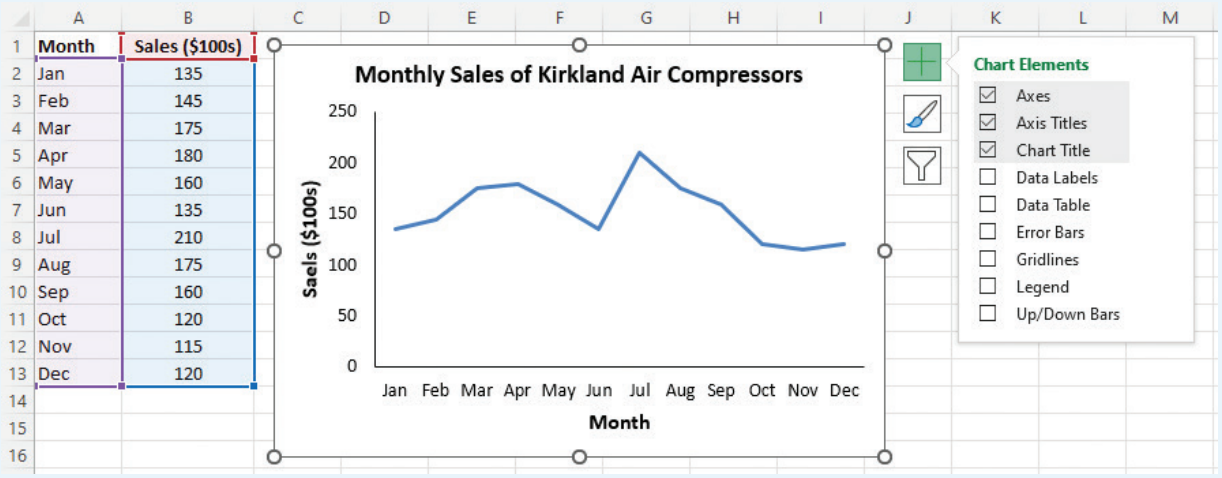
Line charts can also be used to graph multiple lines. Suppose we want to break out Kirkland's sales data by region (North and South), as shown in Table 3.10. We can create a line chart in Excel that shows sales in both regions, as in Figure 3.23 by following similar steps but selecting cells A2:C14 in the file *kirklandregional* before creating the line chart. Figure 3.23 shows an interesting pattern. Sales in both the North and the South regions seemed to follow the same increasing/decreasing pattern until October. Starting in October, sales in the North continued to decrease while sales in the South increased. We would probably want to investigate any changes that occurred in the North region around October. Note that in Figure 3.23 we have used the preattentive attribute of color to clearly differentiate the line corresponding to the North region from the line corresponding to the South region. We have used the colors of blue and orange for these lines because they are easily differentiated and are color-blind friendly.

A special type of line chart is a **sparkline**, which is a minimalist type of line chart that can be placed directly into a cell in Excel. Sparklines contain no axes; they display only the line for the data. Sparklines take up very little space, and they can be effectively used to provide information on overall trends for time series data. Figure 3.24 illustrates the use of sparklines in Excel for the regional sales data. To create a sparkline in Excel:

DATA*file*
kirklandregional

Step 1. Click the **Insert** tab on the Ribbon
Indent the **Line** button  in the **Sparklines** group

Figure 3.22 Line Chart and Excel's Chart Elements Button Options for Monthly Sales Data at Kirkland Industries



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 3.10 Regional Sales Data by Month for Air Compressors at Kirkland Industries

Month	Sales ($100s) North	Sales ($100s) South
Jan	95	40
Feb	100	45
Mar	120	55
Apr	115	65
May	100	60
Jun	85	50
Jul	135	75
Aug	110	65
Sep	100	60
Oct	50	70
Nov	40	75
Dec	40	80

DATA*file*
kirklandregional

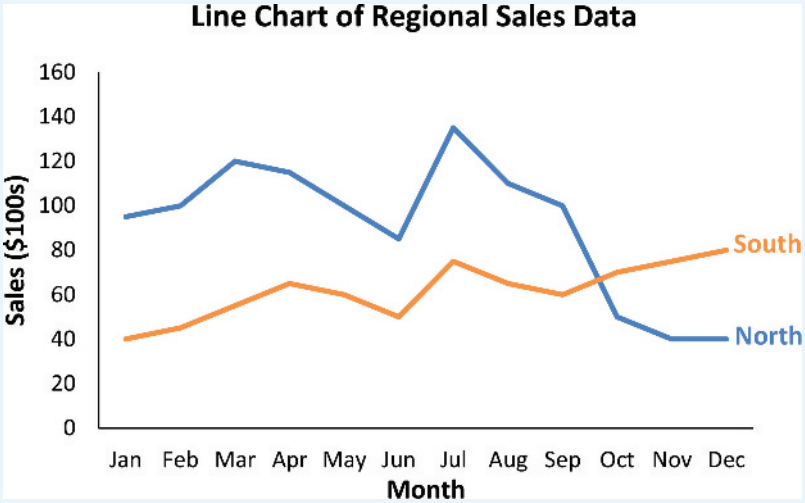
Step 2. When the **Create Sparklines** dialog box appears:
Enter *B3:B14* in the **Data Range:** box
Enter *B15* in the **Location Range:** box
Click **OK**

Step 3. Copy cell B15 to cell C15

The sparklines in cells B15 and C15 do not indicate the magnitude of sales in the North and the South regions, but they do show the overall trend for these data. Sales in the North appear to be decreasing and sales in the South increasing overall. Because sparklines are input directly into the cell in Excel, we can also type text directly into the same cell that will then be overlaid on the sparkline, or we can add shading to the cell, which will appear as the background. Sparklines provide an efficient and simple way to display basic information about a time series.

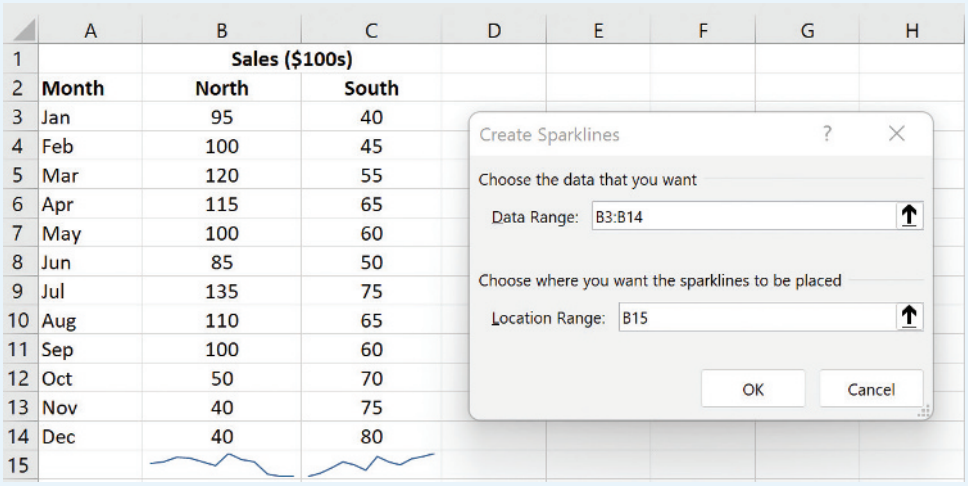
In the line chart in Figure 3.23, we have replaced Excel's default legend with text boxes labeling the lines corresponding to sales in the North and the South. This can often make the chart look cleaner and easier to interpret.

Figure 3.23 Line Chart of Regional Sales Data at Kirkland Industries



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.24 Sparklines for the Regional Sales Data at Kirkland Industries



Bar Charts and Column Charts

DATA*file* accountsmanaged

Bar charts and column charts provide a graphical summary of categorical data using the preattentive attribute of length to convey relative amounts. **Bar charts** use horizontal bars to display the magnitude of the quantitative variable. **Column charts** use vertical bars to display the magnitude of the quantitative variable. Bar and column charts are very helpful in making comparisons between categorical variables. Consider a regional supervisor who wants to examine the number of accounts being handled by each manager. Figure 3.25 shows a bar chart created in Excel displaying these data. To create this bar chart in Excel:

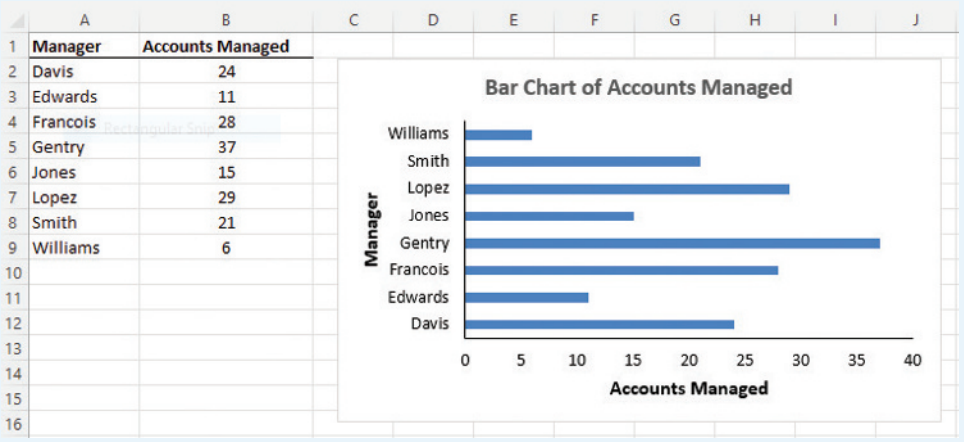
Step 1. Select cells A2:B9
Step 2. Click the **Insert** tab on the Ribbon
Step 3. Click the **Insert Column or Bar Chart** button in the **Charts** group
Step 4. When the list of bar chart subtypes appears:
Click the **Clustered Bar** button in the **2-D Bar** section
Step 5. Select the bar chart that was just created to reveal the **Chart Buttons**
Step 6. Click the **Chart Elements** button
Select the check boxes for **Axes**, **Axis Titles**, and **Chart Title**
Deselect the check box for **Gridlines**
Click on the text box next to the horizontal axis, and replace "Axis Title" with *Accounts Managed*
Click on the text box next to the vertical axis, and replace "Axis Title" with *Manager*
Click on the text box above the chart, and replace "Chart Title" with *Bar Chart of Accounts Managed*

From Figure 3.25, and using the preattentive attribute of length, we can see that Gentry manages the greatest number of accounts and Williams the fewest. We can reduce the cognitive load required of the user and make this bar chart even easier to read by ordering the results by the number of accounts managed. We can do this with the following steps:

Step 1. Select cells A1:B9
Step 2. Right-click any of the cells A1:B9
Select **Sort**
Click **Custom Sort...**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.25 Bar Chart for Accounts Managed Data



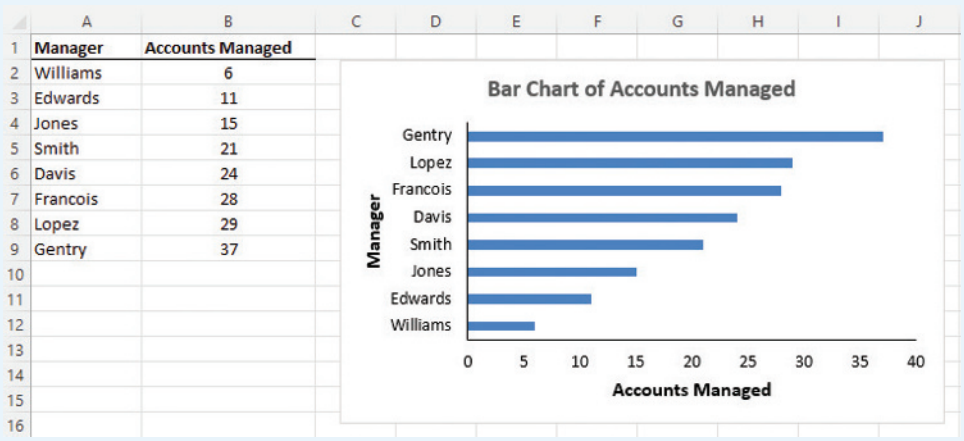
Step 3. When the **Sort** dialog box appears:
Make sure that the check box for **My data has headers** is checked
Select **Accounts Managed** in the **Sort by** box under **Column**
Select **Smallest to Largest** under **Order**
Click **OK**

In the completed bar chart in Excel, shown in Figure 3.26, we can easily compare the relative number of accounts managed for all managers by the length of the respective bars. However, note that it is difficult to interpret from the bar chart exactly how many accounts are assigned to each manager. If this information is necessary, these data are better presented as a table or by adding data labels to the bar chart, as in Figure 3.27, which is created in Excel using the following steps:

*Alternatively, you can add Data Labels by right-clicking on a bar in the chart and selecting **Add Data Labels**.*

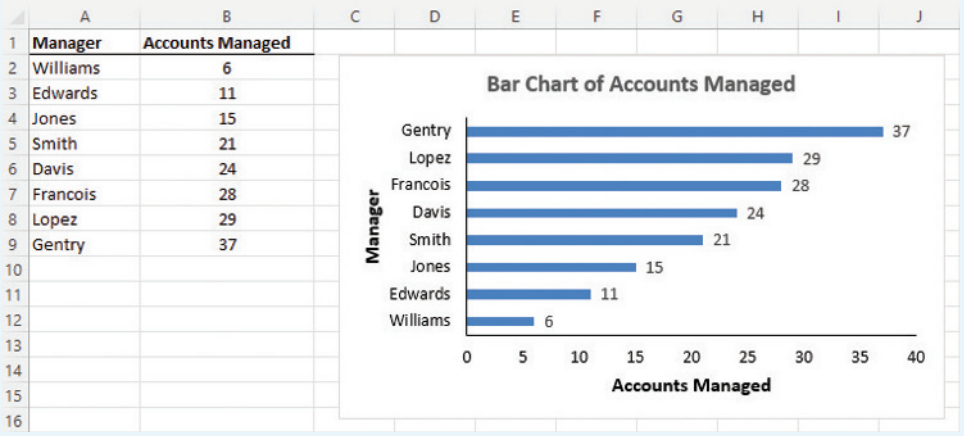
Step 1. Select the chart to reveal the **Chart Buttons**
Step 2. Click the **Chart Elements** button
Select the check box for **Data Labels**

Figure 3.26 Sorted Bar Chart for Accounts Managed Data



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.27 Sorted Bar Chart with Data Labels for Accounts Managed Data



	A	B
1	Manager	Accounts Managed
2	Williams	6
3	Edwards	11
4	Jones	15
5	Smith	21
6	Davis	24
7	Francois	28
8	Lopez	29
9	Gentry	37

This adds labels of the number of accounts managed to the end of each bar so that the reader can easily look up exact values displayed in the bar chart. If this information is important to the user, then the data labels are considered data-ink, and would increase the data-ink ratio of the chart.

Figure 3.28 provides an alternative display for the regional sales data of air compressors for Kirkland Industries. The figure uses a **stacked-column chart** to display the North and the South regional sales data previously shown in a line chart in Figure 3.23. To create the stacked-column chart shown in Figure 3.28, we use the following steps:

*Note that here we have not included the additional steps for formatting the chart in Excel using the **Chart Elements** button, but the steps are similar to those used to create the previous charts.*

Step 1. Select cells A2:C14
Step 2. Click the **Insert** tab on the Ribbon
Step 3. In the **Charts** group, click the **Insert Column or Bar Chart** button
Select **Stacked Column** under **2-D Column**

We could also use a stacked-bar chart to display the same data by using horizontal bars instead of vertical columns. Stacked-column and stacked-bar charts allow the reader to compare the relative values of quantitative variables for the same category in a bar chart. However, these charts use the preattentive attribute of size to differentiate among the different values in the bars (or columns), and because the human eye has difficulty perceiving small differences in areas, it can be difficult to compare the values illustrated within the stacked bars (or columns). As a result, experts often recommend against the use of stacked-column and stacked-bar charts for more than a couple of quantitative variables in each category. An alternative chart for these same data is called a **clustered-column (or clustered-bar) chart**. It is created in Excel following the same steps but selecting **Clustered Column** under the **2-D Column** in step 3. Clustered-column and clustered-bar charts are often superior to stacked-column and stacked-bar charts for comparing quantitative variables, but they can become cluttered for more than a few quantitative variables per category.

Clustered-column (bar) charts are also referred to as side-by-side-column (bar) charts.

An alternative that is often preferred to both stacked and clustered charts, particularly when many quantitative variables need to be displayed, is to use multiple charts. For the regional sales data, we would include two column charts: one for sales in the North and one for sales in the South. For additional regions, we would simply add additional column charts. To facilitate comparisons between the data displayed in each chart, it is important to maintain consistent axes from one chart to another. The categorical variables should be listed in the same order in each chart, and the axis for the quantitative variable should have the same range. For instance, the vertical axis for both North and South sales starts at 0 and ends at 140. This makes it easy to see that, in most months, the North region has greater sales. Figure 3.29 compares the approaches using a stacked-column chart, a clustered-column chart, and multiple column charts for the regional sales data.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.28 Stacked-Column Chart for Regional Sales Data for Kirkland Industries

	A	B	C
1		Sales ($100s)	
2	Month	North	South
3	Jan	95	40
4	Feb	100	45
5	Mar	120	55
6	Apr	115	65
7	May	100	60
8	Jun	85	50
9	Jul	135	75
10	Aug	110	65
11	Sep	100	60
12	Oct	50	70
13	Nov	40	75
14	Dec	40	80
15			
16			

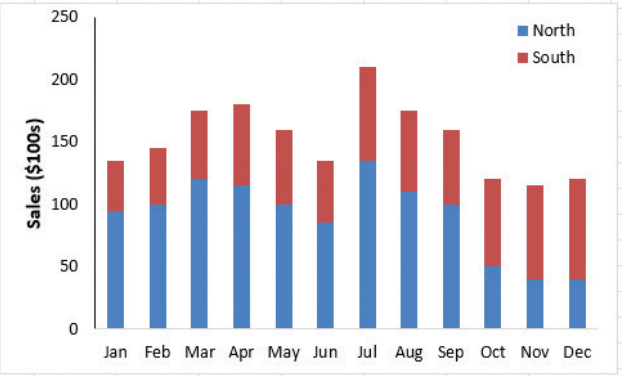
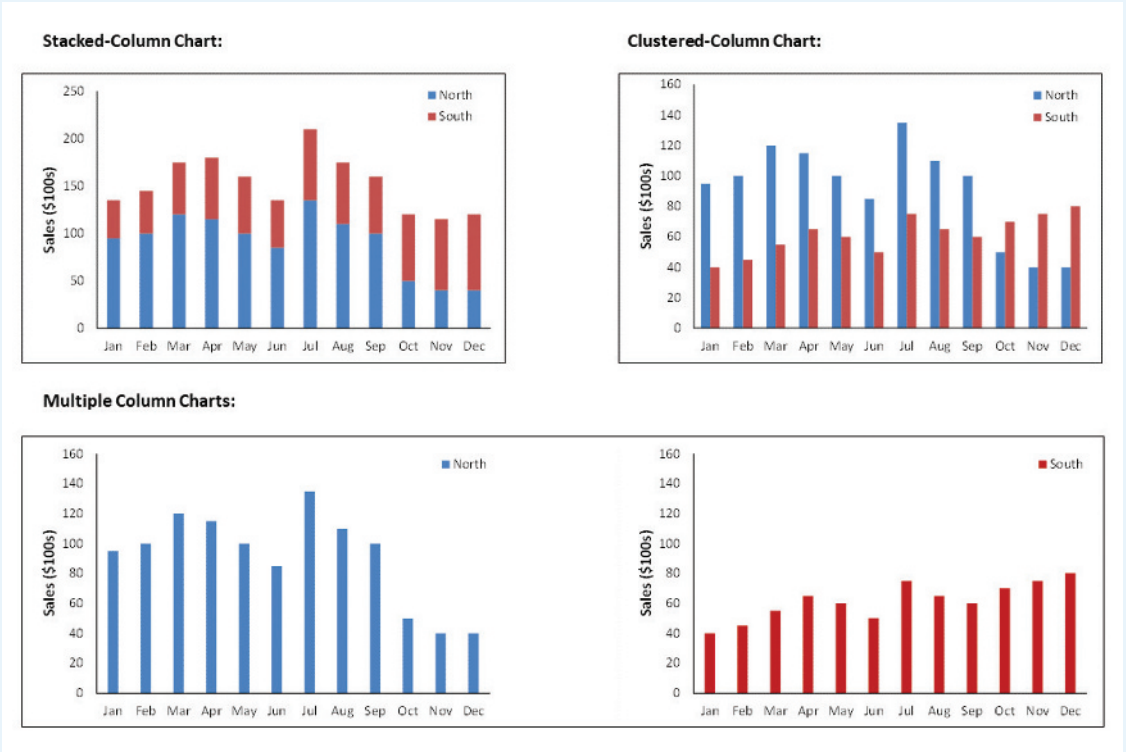


Figure 3.29 shows that multiple column charts require considerably more space than the stacked- and clustered-column charts. However, when comparing many quantitative variables, using multiple charts can often be superior even if each chart must be made smaller. Stacked-column and stacked-bar charts should be used only when comparing a few quantitative variables and when there are large differences in the relative values of the quantitative variables within the category.

Figure 3.29 Comparing Stacked-, Clustered-, and Multiple-Column Charts for the Regional Sales Data for Kirkland Industries



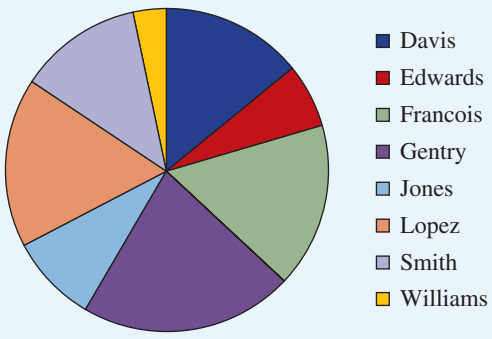
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A Note on Pie Charts and Three-Dimensional Charts

Pie charts are another common form of chart used to compare categorical data. However, pie charts rely on the preattentive attribute of size to convey information to the user. As mentioned previously (see Figure 3.7), most people are not good at differentiating relative amounts using the preattentive attributes of size. Thus, many experts argue that pie charts are inferior to bar charts for comparing data. The pie chart in Figure 3.30 displays the data for the number of accounts managed in Figure 3.25. Visually, it is still relatively easy to see that Gentry has the greatest number of accounts and that Williams has the fewest. However, it is difficult to say whether Lopez or Francois has more accounts because it is difficult to interpret these relative differences in size. Compare Figure 3.30 to Figure 3.26. Making visual comparisons is much easier in the bar chart, which relies on the preattentive attribute of length, than in the pie chart, which uses the preattentive attribute of size. Note that Figure 3.30 also uses the preattentive attribute of color to differentiate among the account managers. This makes it easier to see which portion of the pie chart is associated with each manager, but it still requires considerable cognitive load on the part of the user to match the manager to the correct portion of the pie chart. The number of colors used here could also be distracting, and users who are color-blind may have difficulty differentiating among some of the colors used here. These are all additional reasons why a bar chart is likely to be preferred for this data visualization. Therefore, we recommend against using pie charts in most situations and suggest instead using bar charts for comparing categorical data.

Because of the difficulty in using the preattentive attribute of size to compare relative amounts, many experts also recommend against the use of three-dimensional (3-D) charts in most settings. Excel makes it very easy to create 3-D bar, line, pie, and other types of charts. In most cases, however, the 3-D effect simply adds unnecessary detail that does not help explain the data, and therefore, reduces the data-ink ratio. As an alternative, consider the use of multiple lines on a line chart (instead of adding a z-axis), employing multiple charts, or creating bubble charts in which the size of the bubble can represent the z-axis value. Never use a 3-D chart when a two-dimensional chart will suffice.

Figure 3.30 Pie Chart of Accounts Managed



Additional Visualizations for Multiple Variables: Bubble Chart, Scatter Chart Matrix, and Table Lens

Here we discuss some additional data visualizations that can be used to explore and explain the relationships among multiple variables. Specifically, we introduce the bubble chart, the scatter chart matrix, and the table lens.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A **bubble chart** is a graphical means of visualizing three variables in a two-dimensional graph and is therefore sometimes a preferred alternative to a 3-D graph. Suppose that we want to examine waiting times for passengers at airports along with parking rates and passenger traffic at those same airports. Table 3.11 shows data for a sample of 15 airports. These data are also contained in the file *airportdata*. For each airport, we have the following quantitative variables: average wait time in the nonpriority Transportation Security Authority (TSA) queue measured in minutes, the cheapest on-site daily rate for parking at the airport measured in dollars, and the number of enplanements in a year (the number of passengers who board including transfers) measured in millions of passengers. We can create a bubble chart using Excel to further examine these data by using the steps that follow.

DATA*file*
airportdata

Step 1. Select cells B2:D16

Step 2. Click the **Insert** tab on the Ribbon

Step 3. In the **Charts** group, click **Insert Scatter (X,Y) or Bubble Chart**

In the **Bubble** subgroup, click **Bubble**

Step 4. Select the chart that was just created to reveal the **Chart Buttons**

Step 5. Click the **Chart Elements** button

Select the check boxes for **Axes**, **Axis Titles**, **Chart Title** and **Data Labels**. **Deselect** the check box for **Gridlines**.

Click on the text box under the horizontal axis, and replace "Axis Title" with *TSA Wait Time (minutes)*

Click on the text box next to the vertical axis, and replace "Axis Title" with *Cheapest Parking Rate ($ per day)*

Click on the text box above the chart, and replace "Chart Title" with *Airports Comparison (Bubble Size Corresponds to Annual Enplanements at Each Airport)*

Table 3.11 Sample Data for 15 Airports

Airport Code	TSA wait time (minutes)	Cheapest on-site parking cost ($ per day)	Annual Enplanements (millions of passengers)
ATL	10.30	14.00	49.06
CLT	9.45	7.00	22.19
DEN	8.35	8.00	27.02
AUS	5.30	8.00	5.80
STL	4.80	7.00	6.25
SMF	7.30	10.00	4.60
RDU	6.75	6.50	4.80
EWR	9.90	18.00	18.80
SFO	7.25	18.00	24.00
LAX	6.40	12.00	36.10
SLC	4.50	10.00	10.80
SAN	8.80	15.00	9.70
MSY	4.60	16.00	5.30
IAD	3.00	10.00	10.70
BNA	2.50	12.00	5.60

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Step 6. Double-click on one of the Data Labels in the chart (e.g., the "$14.00" in largest bubble in the chart) to reveal the **Format Data Labels** task pane

Step 7. In the **Format Data Labels** task pane, click the **Label Options** icon and open the **Label Options** area
Under **Label Contains**, select **Value from Cells**

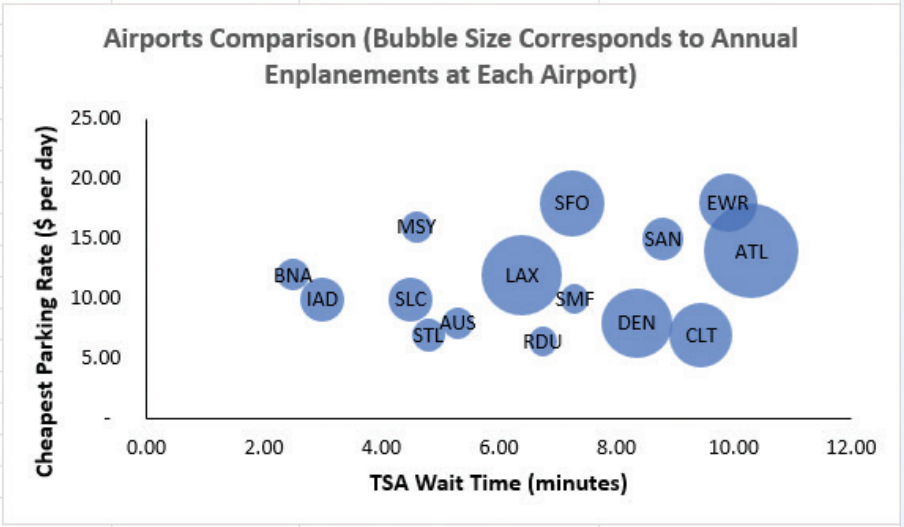
Step 8. When the **Data Label Range** dialog box opens, select cells A2:A16 in the Worksheet
Click **OK**

Step 9. In the **Format Data Labels** task pane, deselect **Y Value** under **Label Contains**, and select **Center** under **Label Position**

The completed bubble chart appears in Figure 3.31. This size of each bubble in Figure 3.31 is proportionate to the number of annual enplanements at each airport.

Figure 3.31 Bubble Chart Comparing Sample of 15 Airports

	A	B	C	D
1	Airport Code	TSA wait time (minutes)	Cheapest on-site parking cost ($ per day)	Annual Enplanements (millions of passengers)
2	ATL	10.30	14.00	49.06
3	CLT	9.45	7.00	22.19
4	DEN	8.35	8.00	27.02
5	AUS	5.30	8.00	5.80
6	STL	4.80	7.00	6.25
7	SMF	7.30	10.00	4.60
8	RDU	6.75	6.50	4.80
9	EWR	9.90	18.00	18.80
10	SFO	7.25	18.00	24.00
11	LAX	6.40	12.00	36.10
12	SLC	4.50	10.00	10.80
13	SAN	8.80	15.00	9.70
14	MSY	4.60	16.00	5.30
15	IAD	3.00	10.00	10.70
16	BNA	2.50	12.00	5.60



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Airports with more passenger traffic appears as larger bubbles, so ATL (Atlanta) is shown to have the most annual enplanements in this sample data. The cheapest parking rate is shown on the vertical axis, and the TSA wait time is shown on the horizontal axis. We see that airports with fewer passengers tend to have lower wait times than those with more passengers. There seems to be less of a relationship between parking rate and number of passengers. Airports with lower wait times do tend to have lower parking rates.

An especially useful chart for displaying multiple variables is the **scatter-chart matrix**. Table 3.12 contains a partial listing of the data for each of New York City's 55 sub-boroughs (a designation of a community within New York City) on median monthly rent, percent college graduates, poverty rate, and mean commute time between home and work. The full data is contained in the file *nyc*. Suppose we want to examine the relationship between these different variables. Figure 3.32 displays a scatter-chart matrix (scatter-plot matrix) for data related to rentals in New York City.

A scatter-chart matrix allows the reader to easily see the relationships among multiple variables. Each scatter chart in the matrix is created in the same manner as for creating a single scatter chart. Each column and row in the scatter-chart matrix corresponds to one categorical variable. For instance, row 1 and column 1 in Figure 3.32 correspond to the median monthly rent variable. Row 2 and column 2 correspond to the percent college graduates variable. Therefore, the scatter chart shown in row 1, column 2 shows the relationship between median monthly rent (on the y-axis) and the percent college graduates (on the x-axis) in New York City sub-boroughs. The scatter chart shown in row 2, column 3 shows the relationship between the percent college graduates (on the y-axis) and poverty rate (on the x-axis).

We demonstrate how to create scatter-chart matrixes in several different software packages in the online appendixes.

Figure 3.32 allows us to infer several interesting findings. Because the points in the scatter chart in row 1, column 2 generally get higher moving from left to right, this tells us that sub-boroughs with higher percentages of college graduates appear to have higher median monthly rents. The scatter chart in row 1, column 3 indicates that sub-boroughs with higher poverty rates

Table 3.12 Rental Data for New York City Sub-Boroughs

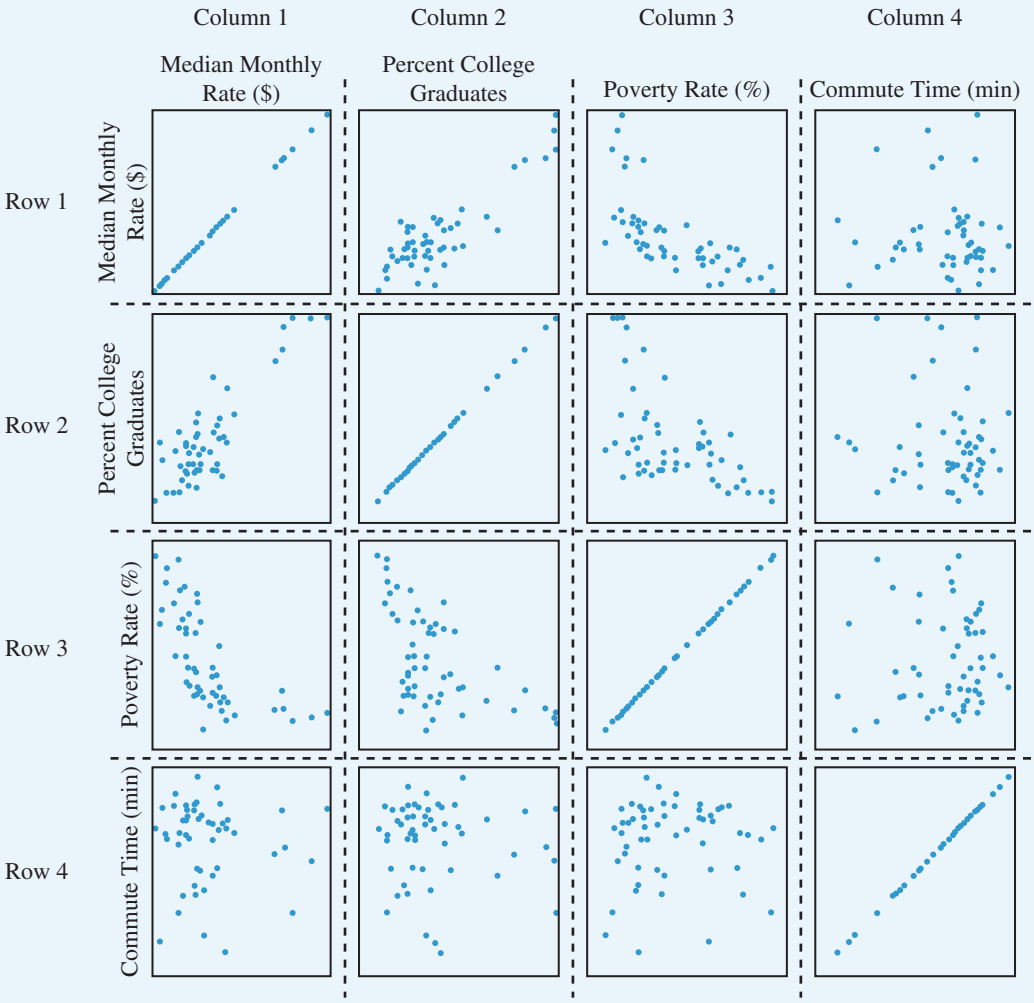
Area (Sub-Borough)	Median Monthly Rent ($)	Percent College Graduates (%)	Poverty Rate (%)	Commute Time (min)
Astoria	1,106	36.8	15.9	35.4
Bay Ridge	1,082	34.3	15.6	41.9
Bayside/Little Neck	1,243	41.3	7.6	40.6
Bedford Stuyvesant	822	21.0	34.2	40.5
Bensonhurst	876	17.7	14.4	44.0
Borough Park	980	26.0	27.6	35.3
Brooklyn Heights/ Fort Greene	1,086	55.3	17.4	34.5
Brownsville/ Ocean Hill	714	11.6	36.0	40.3
Bushwick	945	13.3	33.5	35.5
Central Harlem	665	30.6	27.1	25.0
Chelsea/Clinton/ Midtown	1,624	66.1	12.7	43.7
Coney Island	786	27.2	20.0	46.3
⋮	⋮	⋮	⋮	⋮

DATA*file*
nyc

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The scatter charts along the diagonal in a scatter-chart matrix (e.g., in row 1, column 1 and in row 2, column 2) display the relationship between a variable and itself. Therefore, the points in these scatter charts will always fall along a straight line at a 45-degree angle, as shown in Figure 3.33.

Figure 3.32 Scatter-Chart Matrix for New York City Rental Data



appear to have lower median monthly rents. The data in row 2, column 3 show that sub-boroughs with higher poverty rates tend to have lower percentages of college graduates. The scatter charts in column 4 show that the relationships between the mean travel time and the other variables are not as clear as relationships in other columns.

The scatter-chart matrix is very useful in analyzing relationships among variables. Unfortunately, it is not possible to generate a scatter-chart matrix using standard Excel functions. Each scatter chart must be created individually in Excel using the data from those two variables to be displayed on the chart and then the scatter charts must be arranged manually to create the scatter chart matrix.

A **table lens** is another way to visualize relationships between different pairs of variables which can be created in Excel. In a table lens, a table of data is highlighted with horizontal bars with lengths proportional to the values in each variable's column. A table lens can be a useful visualization tool for wide and tall data sets as the insight on the relationships between variables remains evident even if the display is "zoomed out" to show the table in its entirety or near-entirety.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The following steps demonstrate the construction of a table lens on the quantitative variables in the New York City sub-borough data.

DATA*file* nyc

Step 1. Select cells B2:B56
Step 2. Click the **Home** tab on the Ribbon
Step 3. Click the **Conditional Formatting** button Conditional Formatting in the **Styles** group
From the drop-down menu, select **Data Bars** Data Bars and then **Blue Data Bar** from the **Solid Fill** area

Repeat steps 1–3 for the other three variables, changing the cell range in step 1 to C2:C56, D2:D56, and E2:E56, respectively. This will result in the four variable columns being formatted with horizontal bars proportional to the values of the respective variable. To facilitate comparisons, it is also recommended that each column in the table lens be similar in width.

To visualize relationships between variables with a table lens, we must now sort the data according to one of the quantitative variables. The next steps sort the data according to Median Monthly Rent.

Step 1. Select cells A1:F56
Step 2. Click the **Data** tab on the Ribbon
Step 3. Click the **Sort** button Sort in the **Sort & Filter** group
Step 4. When the **Sort** dialog box appears, select the check box for **My data has headers**
In the **Sort by** row:
Select **Median Monthly Rent ($)** for the **Column** entry
Select **Cell Values** for **Sort on** entry
Select **Largest to Smallest** for **Order** entry
Click **OK**

Figure 3.33 displays the resulting table lens. We interpret this table lens by comparing the large values to small values pattern in the column we sorted (in this case Median Monthly Rent) to the patterns in the columns. Because Percent College Graduates also displays a large values to small values pattern, we can deduce that this variable has a positive association with Median Monthly Rent. Conversely, Poverty Rate displays a small values to large values pattern, so we can deduce that this variable has a negative association with Median Monthly Rent. The Commute Time column displays no pattern, so we can deduce that this variable has no relationship with Median Monthly Rent. By sorting the table on the values of a different variable, the relationships between different pairs of variables can be analyzed.

PivotCharts in Excel

DATA*file* restaurant

To summarize and analyze data with both a chart and a crosstabulation, Excel pairs **PivotCharts** with PivotTables. Using the restaurant data introduced in Table 3.7 and Figure 3.12, we can create a PivotChart by taking the following steps:

Step 1. Click the **Insert** tab on the Ribbon
Step 2. In the **Charts** group, select **PivotChart** PivotChart
Step 3. When the **Create PivotChart** dialog box appears:
Choose **Select a Table or Range**
Enter *A1:D301* in the **Table/Range:** box
Select **New Worksheet** as the location for the PivotTable Report
Click **OK**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.33 Table Lens for New York City Rental Data

	A	B	C	D	E
1	Area (Sub-Borough)	Median Monthly Rent ($)	Percent College Graduates (%)	Poverty Rate (%)	Commute Time (min)
2	Greenwich Village/Financial District	1965	78.3	7.9	44.0
3	Stuyvesant Town/Turtle Bay	1849	78.0	6.9	36.7
4	Upper East Side	1706	78.2	5.9	29.1
5	Upper West Side	1635	74.6	8.8	38.7
6	Chelsea/Clinton/Midtown	1624	66.1	12.7	43.7
7	Park Slope/Carroll Gardens	1566	61.8	8.4	37.5
8	Bayside/Little Neck	1243	41.3	7.6	40.6
9	Rego Park/Forest Hills	1192	51.2	10.4	42.5
10	Queens Village	1184	30.2	6.2	41.3
11	Flushing/Whitestone	1170	32.5	11.4	23.4
12	South Ozone Park/Howard Beach	1148	17.3	8.1	42.0
13	Hillcrest/Fresh Meadows	1138	39.6	12.9	41.6
14	Sunnyside/Woodside	1137	31.8	10.2	44.7
15	Elmhurst/Corona	1121	21.8	22.3	41.1
16	Jackson Heights	1114	19.8	16.0	47.3
17	Ozone Park/Woodhaven	1114	21.7	11.6	35.6
18	Astoria	1106	36.8	15.9	35.4
19	Brooklyn Heights/Fort Greene	1086	55.3	17.4	34.5
20	Bay Ridge	1082	34.3	15.6	41.9
21	Middle Village/Ridgewood	1078	20.1	12.1	39.7
22	Flatlands/Canarsie	1052	26.6	9.3	42.0
23	Mid-Island	999	28.4	11.3	32.5
24	South Shore	999	27.6	4.3	25.9
25	Sunset Park	995	21.8	27.6	43.1
26	Borough Park	980	26.0	27.6	35.3
27	Jamaica	980	19.8	12.7	42.8
28	Riverdale/Kingsbridge	969	41.4	13.4	48.6
29	Williamsburg/Greenpoint	959	33.9	31.7	33.8
30	Williamsbridge/Baychester	950	20.1	16.7	31.9
31	Morningside Heights/Hamilton	949	38.1	25.3	44.8
32	Bushwick	945	13.3	33.5	35.5
33	Sheepshead Bay/Gravesend	945	32.5	11.9	44.4
34	East Flatbush	940	18.4	11.7	33.2
35	Flatbush	931	28.3	25.1	43.9
36	Pelham Parkway	929	22.7	17.4	44.8
37	Throgs Neck/Co-op City	892	22.3	13.6	39.7
38	East New York/Starrett City	890	13.7	29.2	43.9
39	North Shore	889	26.5	17.4	43.0
40	Bensonhurst	876	17.7	14.4	44.0
41	Rockaways	875	22.2	20.1	40.4
42	South Crown Heights	874	19.9	19.7	44.6
43	Washington Heights/Inwood	874	28.8	25.9	42.5
44	North Crown Heights/Prospect	872	29.7	24.8	42.8
45	Kingsbridge Heights/Moshulu	860	16.0	35.0	31.6
46	Soundview/Parkchester	848	15.8	27.9	42.4
47	Bedford Stuyvesant	822	21.0	34.2	40.5
48	Lower East Side/Chinatown	821	34.2	25.9	39.0
49	University Heights/Fordham	813	11.6	40.9	29.2
50	Coney Island	786	27.2	20.0	46.3
51	Highbridge/South Concourse	781	11.2	31.4	44.6
52	Morrisania/Belmont	727	11.3	38.9	39.7
53	Brownsville/Ocean Hill	714	11.6	36.0	40.3
54	East Harlem	677	23.7	30.0	44.2
55	Central Harlem	665	30.6	27.1	25.0
56	Mott Haven/Hunts Point	625	8.3	41.6	41.3

Step 4. In the **PivotChart Fields** area, under **Choose fields to add to report:**
Drag the **Quality Rating** field to the **Axis (Categories)** area
Drag the **Meal Price ($)** field to the **Legend (Series)** area
Drag the **Wait Time (min)** field to the **Values** area

Step 5. Click on **Sum of Wait Time (min)** in the **Values** area

Step 6. Select **Value Field Settings…** from the list of options that appear

Step 7. When the **Value Field Settings** dialog box appears:
Under **Summarize value field by**, select **Average**
Click **Number Format**

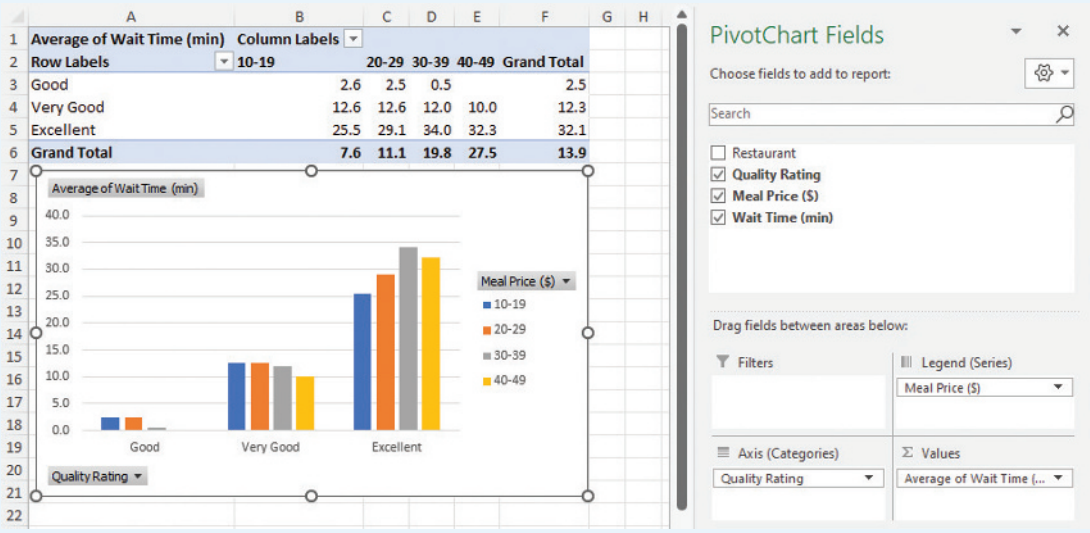
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In the **Category:** box, select **Number**
Enter *1* for **Decimal places:**
Click **OK**
When the **Value Field Settings** dialog box reappears, click **OK**

Step 8. Right-click in cell B2 or any cell containing a meal price column label
Step 9. Select **Group...** from the list of options that appears
Step 10. When the **Grouping** dialog box appears:
Enter *10* in the **Starting at:** box
Enter *49* in the **Ending at:** box
Enter *10* in the **By:** box
Click **OK**
Step 11. Right-click on "Excellent" in cell A3
Step 12. Select **Move** and click **Move "Excellent" to End**

The completed PivotTable and PivotChart appear in Figure 3.34. The PivotChart is a clustered-column chart whose column heights correspond to the average wait times and are clustered into the categorical groupings of Good, Very Good, and Excellent. The columns are different colors to differentiate the wait times at restaurants in the various meal price ranges. Figure 3.34 shows that Excellent restaurants have longer wait times than Good and Very Good restaurants. We also see that Excellent restaurants in the price range of $30–$39 have the longest wait times. The PivotChart displays the same information as that of the PivotTable in Figure 3.18, but the column chart used here makes it easier to compare the restaurants based on quality rating and meal price. Like PivotTables in Excel, PivotCharts in Excel are interactive. You can use the arrows on the axes and legend labels to change the categorical data being displayed. For example, you can click on the **Quality Rating** horizontal axis label (see Figure 3.34) and choose to look at only Very Good and Excellent restaurants, or you can click on the **Meal Price ($)** legend label and choose to view only certain meal price categories.

Figure 3.34 PivotTable and PivotChart for the Restaurant Data



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notes + Comments

1. Excel assumes that line charts will be used to graph only time series data. The Line Chart tool in Excel is the most intuitive for creating charts that include text entries for the horizontal axis (e.g., the month labels of Jan, Feb, Mar, etc. for the monthly sales data in Figure 3.21). When the horizontal axis represents numerical values (1, 2, 3, etc.), then it is easiest to go to the **Charts** group under the **Insert** tab in the Ribbon, click the **Insert Scatter (X,Y) or Bubble Chart** button, and then select the **Scatter with Straight Lines and Markers** button.
2. Histograms and boxplots (discussed in Chapter 2 in relation to analyzing distributions) are other effective data-visualization tools for summarizing the distribution of data.
3. Because PivotCharts are interactive, they can be used to create Data Dashboards in Excel.

3.4 Specialized Data Visualizations

In this chapter, we have presented some of the most basic ideas for using data visualization effectively both to analyze data and to communicate data analysis to others. The charts discussed so far are those most commonly used and will suffice for most data-visualization needs. However, many additional concepts, charts, and tools can be used to improve your data-visualization techniques. In this section, we briefly mention some of these specialized charts including heat maps, treemaps, waterfall charts, stock charts, and parallel-coordinates plots.

Heat Maps

A **heat map** is a two-dimensional graphical representation of data that uses different shades of color to indicate magnitude. Figure 3.35 shows a heat map indicating the magnitude of changes for a metric called same-store sales, which are commonly used in the retail industry to measure trends in sales. The cells shaded red in Figure 3.35 indicate declining same-store sales for the month, and cells shaded blue indicate increasing same-store sales for the month. Column N in Figure 3.35 also contains sparklines for the same-store sales data.

DATA*file*
samestore_sales

Figure 3.35 can be created in Excel by following these steps:

Step 1. Select cells B2:M17
Step 2. Click the **Home** tab on the Ribbon

Figure 3.35 Heat Map and Sparklines for Same-Store Sales Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	SPARKLINES
2	St. Louis	-2%	-1%	-1%	0%	2%	4%	3%	5%	6%	7%	8%	8%	
3	Phoenix	5%	4%	4%	2%	2%	-2%	-5%	-8%	-6%	-5%	-7%	-8%	
4	Albany	-5%	-6%	-4%	-5%	-2%	-5%	-5%	-3%	-1%	-2%	-1%	-2%	
5	Austin	16%	15%	15%	16%	18%	17%	14%	15%	16%	19%	18%	16%	
6	Cincinnati	-9%	-6%	-7%	-3%	3%	6%	8%	11%	10%	11%	13%	11%	
7	San Francisco	2%	4%	5%	8%	4%	2%	4%	3%	1%	-1%	1%	2%	
8	Seattle	7%	7%	8%	7%	5%	4%	2%	0%	-2%	-4%	-6%	-5%	
9	Chicago	5%	3%	2%	6%	8%	7%	8%	5%	8%	10%	9%	8%	
10	Atlanta	12%	14%	13%	17%	12%	11%	8%	7%	7%	8%	5%	3%	
11	Miami	2%	3%	0%	1%	-1%	-4%	-6%	-8%	-11%	-13%	-11%	-10%	
12	Minneapolis	-6%	-6%	-8%	-5%	-6%	-5%	-5%	-7%	-5%	-2%	-1%	-2%	
13	Denver	5%	4%	1%	1%	2%	3%	1%	-1%	0%	1%	2%	3%	
14	Salt Lake City	7%	7%	7%	13%	12%	8%	5%	9%	10%	9%	7%	6%	
15	Raleigh	4%	2%	0%	5%	4%	3%	5%	5%	9%	11%	8%	6%	
16	Boston	-5%	-5%	-3%	4%	-5%	-4%	-3%	-1%	1%	2%	3%	5%	
17	Pittsburgh	-6%	-6%	-4%	-5%	-3%	-3%	-1%	-2%	-2%	-1%	-2%	-1%	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Step 3. Click **Conditional Formatting** in the **Styles** group
Select **Color Scales** and click on **Blue–White–Red Color Scale**

To add the sparklines in column N, we use the following steps:

Step 4. Select cell N2
Step 5. Click the **Insert** tab on the Ribbon
Step 6. Click **Line** in the **Sparklines** group
Step 7. When the **Create Sparklines** dialog box appears:
Enter *B2:M2* in the **Data Range:** box
Enter *N2* in the **Location Range:** box
Click **OK**
Step 8. Copy cell N2 to N3:N17

The heat map in Figure 3.35 helps the reader to easily identify trends and patterns. We can see that Austin has had positive increases throughout the year, while Pittsburgh has had consistently negative same-store sales results. Same-store sales at Cincinnati started the year negative but then became increasingly positive after May. In addition, we can differentiate between strong positive increases in Austin and less substantial positive increases in Chicago by means of color shadings. A sales manager could use the heat map in Figure 3.35 to identify stores that may require intervention and stores that may be used as models. Heat maps can be used effectively to convey data over different areas, across time, or both, as seen here.

*Both the heat map and the sparklines described here can also be created using the **Quick Analysis** button. To display this button, select cells B2:M17. The **Quick Analysis** button will appear at the bottom right of the selected cells. Click the button to display options for heat maps, sparklines, and other data-analysis tools.*

Because heat maps depend strongly on the use of color to convey information, one must be careful to make sure that the colors can be easily differentiated and that they do not become overwhelming. To avoid problems with interpreting differences in color, we can add sparklines as shown in column N of Figure 3.35. The sparklines clearly show the overall trend (increasing or decreasing) for each location. However, we cannot gauge differences in the magnitudes of increases and decreases among locations using sparklines. The combination of a heat map and sparklines here is a particularly effective way to show both trend and magnitude.

Treemaps

A **treemap** is a chart that uses the size, color, and arrangement of rectangles to display the magnitudes of a quantitative variable for different categories, each of which are further decomposed into subcategories. The size of each rectangle represents the magnitude of the quantitative variable within a category/subcategory. The color of the rectangle represents the category and all subcategories of a category are arranged together.

Categorical data that is further decomposed into subcategories is called hierarchical data. **Hierarchical data** can be represented with a tree-like structure, where the branches of the tree lead to categories and subcategories. As an example, let us consider the top ten brand values shown in Table 3.13 (source: *Forbes.com*). Each observation consists of an industry, a brand within an industry, and the value of the brand.

Figure 3.36 shows how these data have a hierarchical or tree structure. The base of the tree is the top ten brand values. The category is the industry of each company, the subcategory is the brand name, and the value of the brand is the quantitative variable.

Figure 3.37 is an example of a treemap created in Excel for the brand values data. The following steps are used to create a treemap in Excel using the data in the file *brandvalues*.

DATA*file*
brandvalues

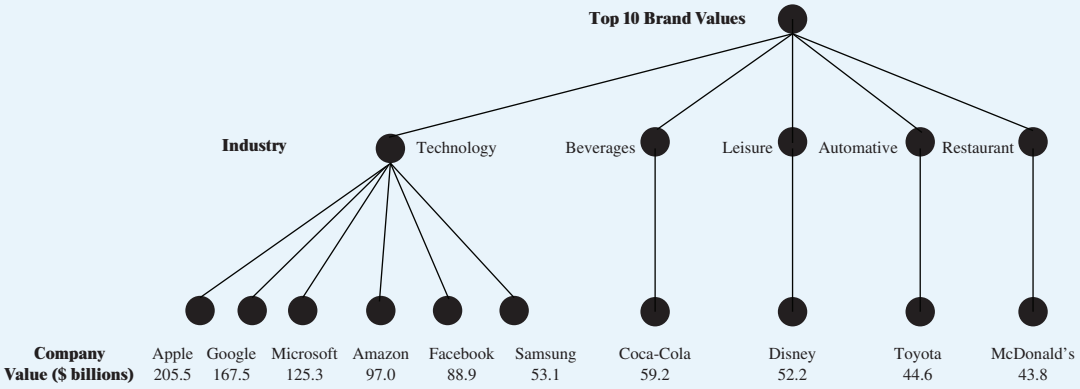
Step 1. Select cells A1:C11
Step 2. Sort the data by Industry by using the following steps:
Click **Data** on the Ribbon
Click the **Sort** button in the **Sort & Filter** group
When the **Sort** dialog box appears:
Select **Industry** for **Sort By**
Select **Cell Values** for **Sort On**
Select **A to Z** for **Order**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 3.13 Top Ten Brand Values Data

Industry	Brand	Value ($ billions)
Technology	Apple	205.5
Technology	Google	167.7
Technology	Microsoft	125.3
Technology	Amazon	97.0
Technology	Facebook	88.9
Beverages	Coca-Cola	59.2
Technology	Samsung	53.1
Leisure	Disney	52.2
Automotive	Toyota	44.6
Restaurants	McDonald's	43.8

Figure 3.36 The Hierarchical Tree Structure of the Top Ten Brand Values Data



Step 3. Click **Insert** on the Ribbon.
Click the **Insert Hierarchy Chart** button in the **Charts** group and select **Treemap**

The following step displays the brand values in the treemap.

Step 4. Right-click on any brand label in the treemap chart (for example, "Apple")
Select **Format Data Labels…**
When the **Format Data Labels** task pane appears:
Select **Label Options** and select **Value** under **Label Contains**

The colors in the treemap shown in 3.37 correspond to industries. Each of the rectangles represents a brand and the size of the rectangles indicates the size of the brand's value. Because it is difficult to compare relative values based on only the preattentive attribute of size, it is a good idea to add the numerical brand values to the treemap as done in the preceding step 4. We see that the technology industry has six brands in the top ten. Apple, Google, and Microsoft are the three highest brand values.

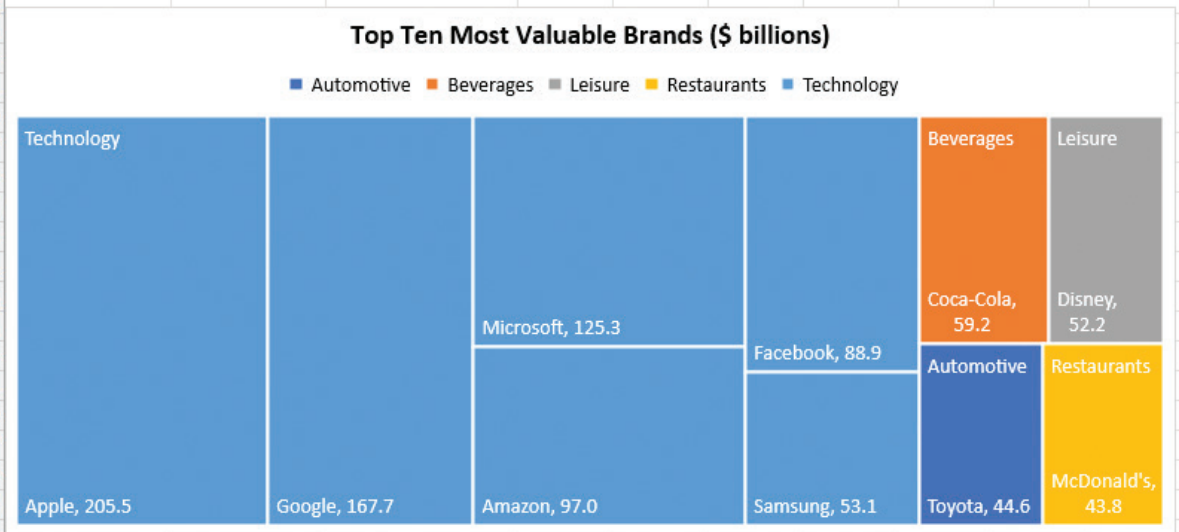
Waterfall Charts

A **waterfall chart** is a visual display that shows the cumulative effect of positive and negative changes on a variable of interest. The changes in a variable of interest are reported

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.37 Treemap Created in Excel for Top Ten Brand Values Data

	Industry	Brand	Value ($ billions)
2	Automotive	Toyota	44.6
3	Beverages	Coca-Cola	59.2
4	Leisure	Disney	52.2
5	Restaurants	McDonald's	43.8
6	Technology	Apple	205.5
7	Technology	Google	167.7
8	Technology	Microsoft	125.3
9	Technology	Amazon	97.0
10	Technology	Facebook	88.9
11	Technology	Samsung	53.1



for a series of categories (such as time periods) and the magnitude of each change is represented by a column anchored at the cumulative height of the changes in the preceding categories.

Consider the costs and revenues data for Gossamer Industries introduced in Table 3.4. Gross Profit is defined as revenues minus variable costs. Table 3.14 shows the costs, revenues and gross profits by month for Gossamer Industries.

The following steps are used to create the waterfall chart of gross profit using the data in file *gossamer_ profit* as shown in Figure 3.38.

DATA*file*
gossamer_profit

Step 1. Select cells A2:H2. Hold down the control key (**Ctrl**) and also select cells A5:H5

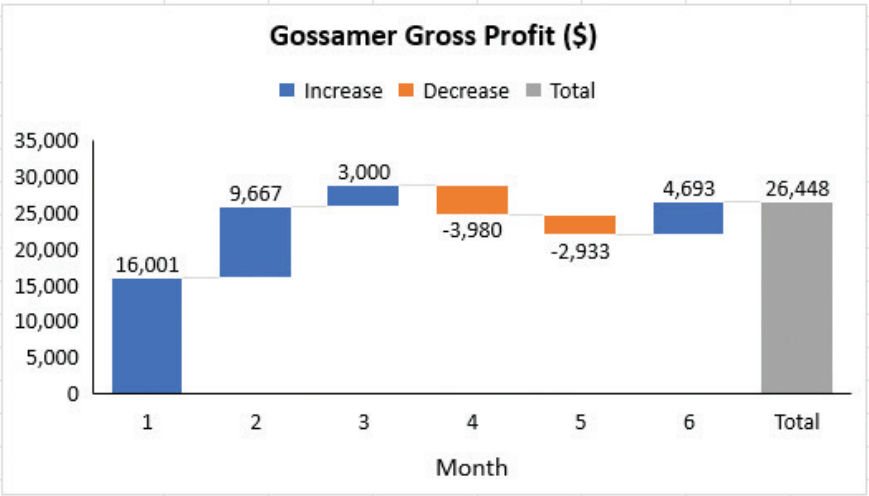
Table 3.14 Costs, Revenues, and Gross Profits Data for Gossamer Industries

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,125	67,125	48,178	51,785	55,678	353,015
Gross Profit ($)	16,001	9,667	3,000	−3,980	−2,933	4,693	26,448

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.38 Waterfall Chart for the Gossamer Gross Profit Data

	A	B	C	D	E	F	G	H
1		Month						
2		1	2	3	4	5	6	Total
3	Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
4	Revenues ($)	64,124	66,125	67,125	48,178	51,785	55,678	353,015
5	Gross Profit ($)	16,001	9,667	3,000	-3,980	-2,933	4,693	26,448



Step 2. Click the **Insert** tab on the Ribbon

Step 3. Click the **Insert Waterfall, Funnel, Stock, Surface or Radar Chart** button in the **Charts** group

When the list of subtypes appears, click **Waterfall**

*In step 4, make sure that you have chosen only the single column in the chart for "Total" to display the **Format Data Point** task pane rather than the Format Data Series task pane.*

In the initial chart, the Total values are treated as if they are another monthly value. The following steps will make the total appear as in Figure 3.38.

Step 4. Double-click on the column for **Total** in the chart to open the **Format Data Point** task pane

Step 5. When the **Format Data Point** task pane appears:

Click the **Series Options** button

Select the check box for **Set as total**

Figure 3.38 shows the gross profit by month, with blue indicating a positive gross profit and orange indicating a negative gross profit. The upper or lower level of the bar indicates the cumulative level of gross profit. For positive changes, the upper level of the bar is the cumulative level, and for negative changes, the lower end of the bar is the cumulative level. Here we see that cumulative level of gross profit rises from January to March (months 1 to 3), drops in April and May (months 4 and 5) and then increases in June (month 6) to the cumulative gross profit of $26,448 for the six-month period.

Stock Charts

A **stock chart** is a graphical display of stock prices over time. Let us consider the stock price data for telecommunication company Verizon Communications shown in Table 3.15. The data in Table 3.15 lists, for five trading days in April: the date, the high price (highest

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
verizon

Table 3.15 Stock Price for Verizon Communications

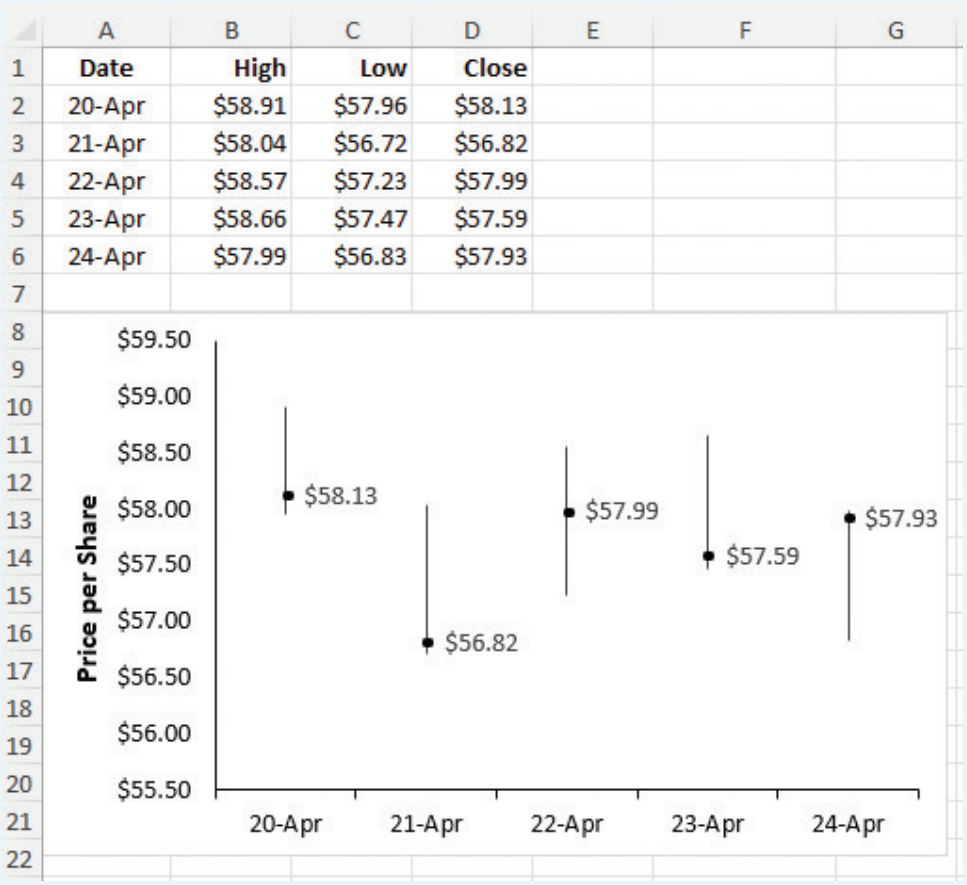
Date	High	Low	Close
20-Apr	$58.91	$57.96	$58.13
21-Apr	$58.04	$56.72	$56.82
22-Apr	$58.57	$57.23	$57.99
23-Apr	$58.66	$57.47	$57.59
24-Apr	$57.99	$56.83	$57.93

price per share observed during the trading day), the low price (the lowest price per share observed during the trading day), and the closing price (the price per share at the end of the trading day).

Excel provides four different types of stock charts. We illustrate the simplest one here, the high-low-close stock chart. A **high-low-close stock chart** is a chart that shows the high value, low value, and closing value of the price of a share of a stock at several points in time. The difference between the highest and lowest share prices for each point in time is represented by a vertical bar, and the closing share price by a marker on the bar.

The following steps are used to create Figure 3.39, the high-low-close stock chart for the Verizon stock price data.

Figure 3.39 A High-Low-Close Stock Chart for Verizon Wireless



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Step 1. Select cells A1:D6
Step 2. Click the **Insert** tab on the Ribbon
Step 3. Click the **Insert Waterfall, Funnel, Stock, Surface or Radar Chart** button in the **Charts** group
When the list of subtypes appears, click **High-Low-Close**

The following steps add the closing price labels and markers.

Step 4. Click the **Chart Elements** button and select **Data Labels**

Step 4 places three sets of labels on each vertical bar (highest, closing, and lowest price per share). The following steps clean up the display.

Step 5. Click any of the high price per share labels (for example, "$58.91" for 20-Apr) and press the **Delete** key.
Do the same for the low price per share labels (for example, "$57.23" for 22-Apr)
Step 6. On one of the vertical lines, click a data point directly next to one of the closing price labels to open the **Format Data Series** task pane
Step 7. When the **Format Data Series** task pane appears:
Click the **Fill & Line** button, then click **Marker**
Under **Fill**, select **Solid fill**
Under **Border**, select **Solid line** and for **Width**, select **3 pt**

As shown in Figure 3.39, the closing prices per share over the five days are given. We see that on April 20, 21, and 23, the price closed near the low end of the trading price range. On April 24, the closing price was near the highest price of the day. On April 22, the closing price was near the middle of the trading price range.

Parallel-Coordinates Chart

A helpful chart for examining data with more than two variables is the **parallel-coordinates plot**, which includes a different vertical axis for each variable. Each observation in the data set is represented by drawing a line on the parallel-coordinates plot connecting each vertical axis. The height of the line on each vertical axis represents the value taken by that observation for the variable corresponding to the vertical axis. For instance, Figure 3.40 displays a parallel coordinates plot for a sample of Major League Baseball players. The figure contains data for 10 players who play first base (1B) and 10 players who play second base (2B). For each player, the leftmost vertical axis plots their total number of home runs (HR). The center vertical axis plots the player's total number of stolen bases (SB), and the rightmost vertical axis plots the player's batting average. Various colors differentiate 1B players from 2B players (1B players are in blue and 2B players are in red).

We can make several interesting statements upon examining Figure 3.40. The sample of 1B players tend to hit lots of HR but have very few SB. Conversely, the sample of 2B players steal more bases but generally have fewer HR, although some 2B players have many HR and many SB. Finally, 1B players tend to have higher batting averages (AVG) than 2B players. We may infer from Figure 3.40 that the traits of 1B players may be different from those of 2B players. In general, this statement is true. Players at 1B tend to be offensive stars who hit for power and average, whereas players at 2B are often faster and more agile in order to handle the defensive responsibilities of the position (traits that are not common in strong HR hitters). Parallel-coordinates plots, in which you can differentiate categorical variable values using color as in Figure 3.40, can be very helpful in identifying common traits across multiple dimensions.

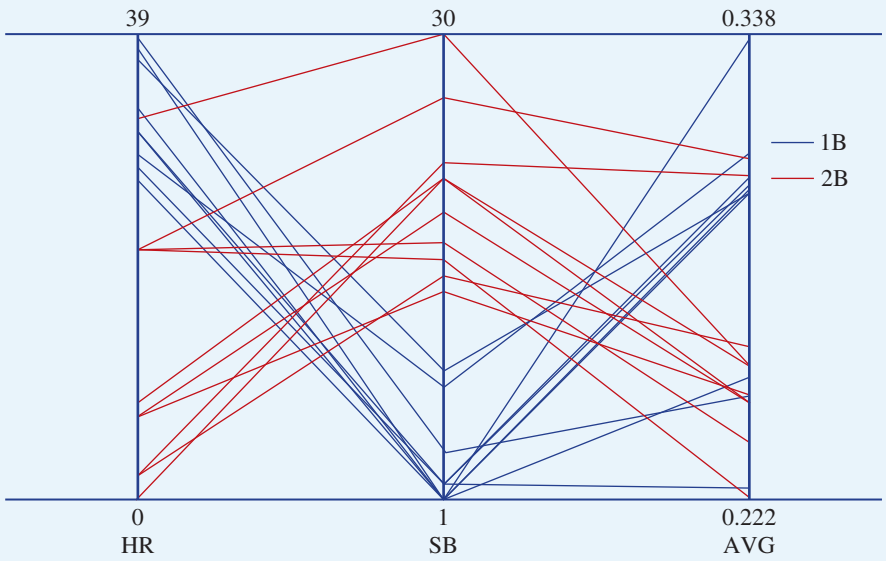
3.5 Visualizing Geospatial Data

Systems that combine the use of geospatial data with geographic visualizations are also known as geographic information systems (GIS).

Geospatial data are data that include information on the geographic location of each record. When dealing with a geospatial data, displaying them on a map can often help to discover patterns that would have been otherwise difficult to interpret. In this section,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.40 Parallel Coordinates Plot for Baseball Data



DATA*file*
baseballdata

we consider two general types of geographic visualizations and discuss their use in data exploration.

Choropleth Maps

A **choropleth map** is a geographic visualization that uses shades of a color, different colors, or symbols to indicate the values of a quantitative or categorical variable associated with a geographic region or area. A familiar example of choropleth map is a weather map, such as the one in Figure 3.41. In Figure 3.41, the preattentive attribute of color is used to depict the daily high temperature, with warmer colors (on the red end of the spectrum) representing higher temperatures and cooler colors (on the purple end of the spectrum) representing lower temperatures.

While choropleth maps may provide a good visual display of changes in a variable between geographic areas, they can also be misleading. If the location data are not granular enough so that the value of the displayed variable is relatively uniform over the respective areas over which it is displayed, then the values of the variable within regions and between regions may be misrepresented. The choropleth map may mask substantial variation of the variable within an area of the same color shading. Further, the choropleth map may suggest abrupt changes in the variable between region boundaries while the actual changes across boundaries may be more gradual.

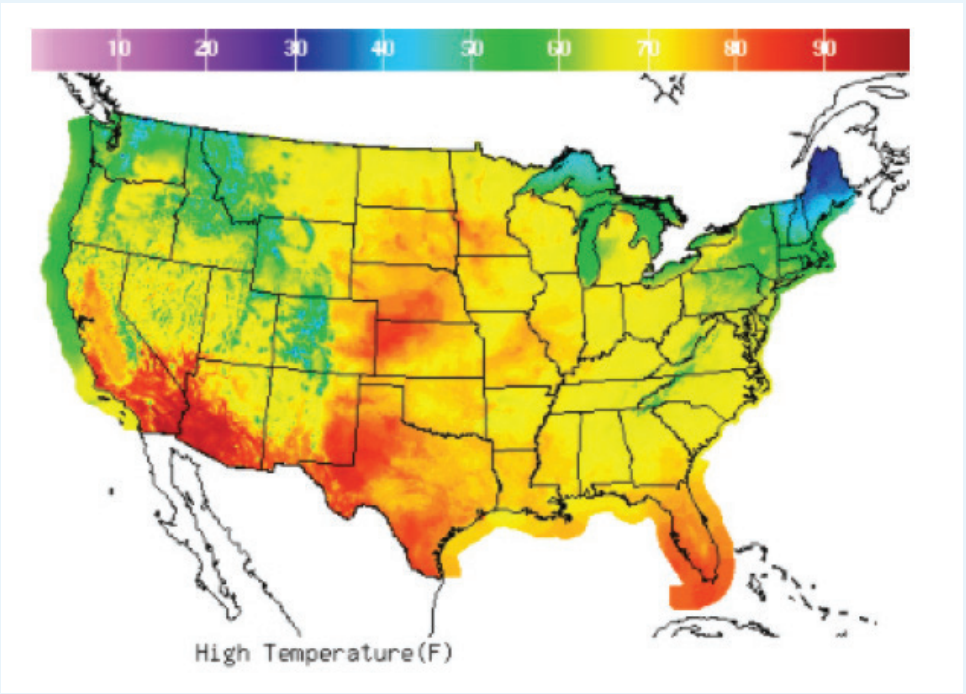
Choropleth maps are the most reliable when the variable displayed is relatively constant within the different locations to be colored. If this is not the case and a choropleth map is desired, the likelihood of the map to convey erroneous insights is mitigated when (1) variable measures are density based (quantity divided by land area or population) or (2) the colored regions are roughly equal-sized so there are no regions that are visually distracting.

DATA*file*
income_state

For data with variables representing geographical regions (e.g., countries, states, counties, postal codes), Excel has mapping functionality that will create a choropleth map. For example, the data in Table 3.16 contains the median income for each state in the United States (also in file *income_state*). The following steps use the file *income_state* to create a choropleth map using shading to denote the median income in each state.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.41 Weather Map as an Example of a Choropleth Map



Excel's Filled Map tool is powered by the Bing search engine. Excel may ask you to accept data being sent from Bing.

Step 1. Select cells A1:B51
Step 2. Click the **Insert** tab on the Ribbon
Step 3. In the **Charts** group, click the **Maps** button 

Step 4. Select **Filled Map** 

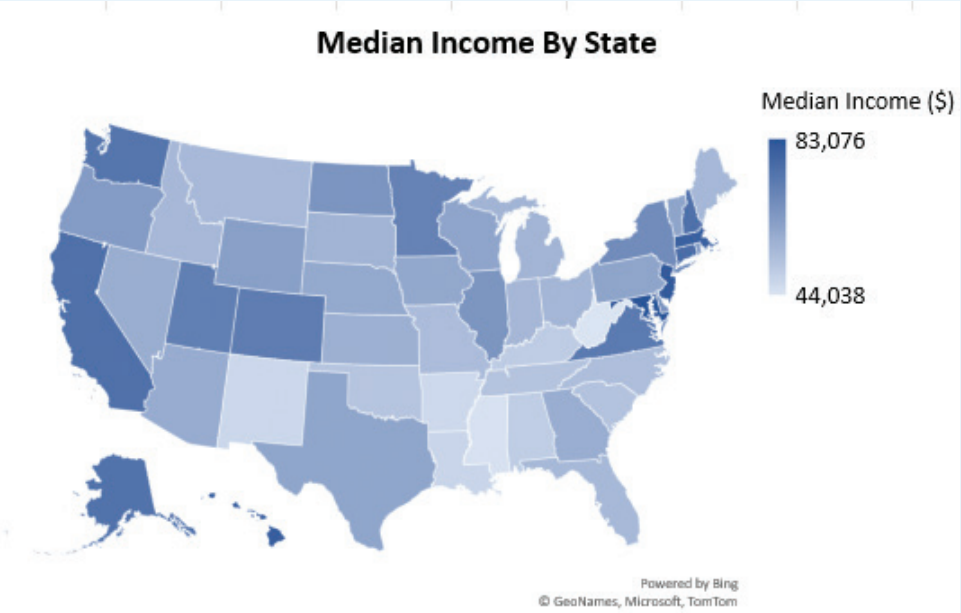
The preceding steps produce the choropleth map in Figure 3.42. As Figure 3.42 demonstrates, choropleth maps are typically better for displaying relative comparisons of magnitude than conveying absolute measures of magnitude. The value of median income for each state

Table 3.16 Median Income Data by State

State	Median Income ($)	State	Median Income ($)	State	Median Income ($)	State	Median Income ($)	State	Median Income ($)
Alabama	49,881	Hawaii	80,527	Massachusetts	79,817	New Mexico	47,405	South Dakota	57,041
Alaska	74,912	Idaho	55,524	Michigan	56,640	New York	67,648	Tennessee	52,366
Arizona	59,079	Illinois	65,063	Minnesota	70,295	North Carolina	53,922	Texas	60,632
Arkansas	47,094	Indiana	55,725	Mississippi	44,740	North Dakota	65,250	Utah	71,381
California	75,250	Iowa	60,071	Missouri	54,412	Ohio	56,155	Vermont	60,917
Colorado	71,949	Kansas	58,150	Montana	55,248	Oklahoma	51,914	Virginia	72,600
Connecticut	76,366	Kentucky	50,189	Nebraska	59,724	Oregon	63,168	Washington	74,023
Delaware	65,467	Louisiana	48,021	Nevada	58,740	Pennsylvania	60,891	West Virginia	44,038
Florida	55,433	Maine	55,579	New Hampshire	75,322	Rhode Island	63,787	Wisconsin	60,779
Georgia	58,634	Maryland	83,076	New Jersey	81,777	South Carolina	52,449	Wyoming	62,259

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.42 Choropleth Map of United States Displaying Median Income by State



is difficult to estimate from Figure 3.42, but it is relatively easy to conclude that northeastern states have high median incomes relative to the rest of the country and many southern states have low median incomes. Indeed, the strength of a choropleth map is the identification of the high-level characteristics of a variable with respect to geographic positioning.

We also observe in Figure 3.42 that the relative size of the states plays a role in the audience's perception. For instance, Rhode Island is barely visible while larger states dominate the audience's field of vision. Another weakness of Figure 3.42 is that it masks the income distributions within each state.

The file *income_county* contains the median income data at the county level in the United States (Figure 3.43). The following steps use the file *income_county* to create a choropleth map using shading to denote the median income in the counties within each state of the United States.

DATA *file*
income_county

Step 1. Select cells A1:C3148
Step 2. Click the **Insert** tab on the Ribbon
Step 3. In the **Charts** group, click the **Maps** button
Step 4. Select **Filled Map**

The preceding steps produce the choropleth map in Figure 3.43.

The choice of choropleth map (state-level in Figure 3.42 or county-level in Figure 3.43) depends on the insight intended to be delivered using the visualization. These two maps could also be used in tandem by first showing the state-level measure and then selecting one or more states to "zoom" in on and show the county-level measure.

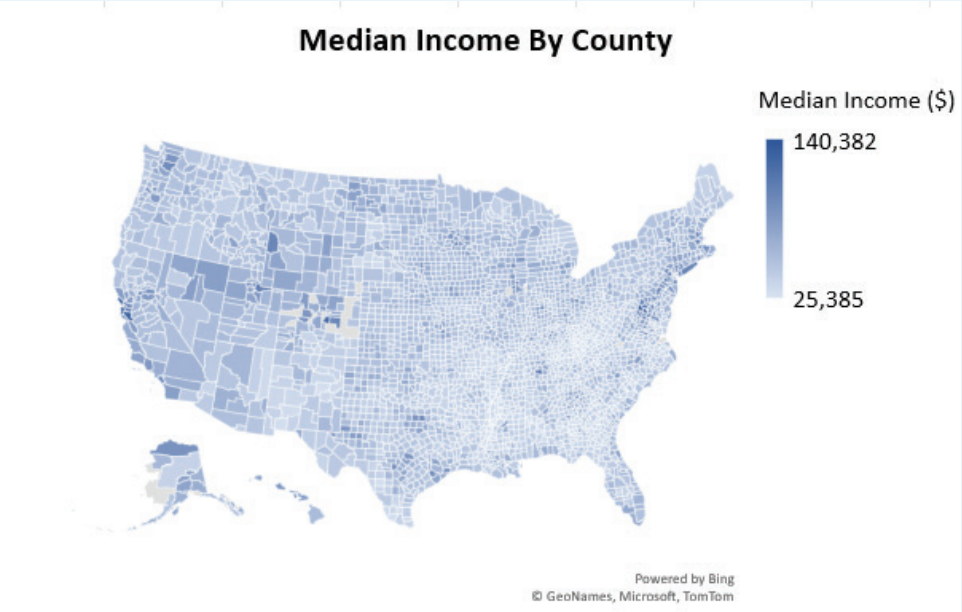
Cartograms

Excel does not have the functionality to automatically generate cartograms.

Another type of geographic display is a cartogram. A **cartogram** is a map-like diagram that uses geographic positioning but purposefully represents map regions in a manner that does not necessarily correspond to land area. For example, Figure 3.44 contains a cartogram of the United States in which the area of each state is based on its population. A cartogram

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.43 Choropleth Map of United States Displaying Median Income by County

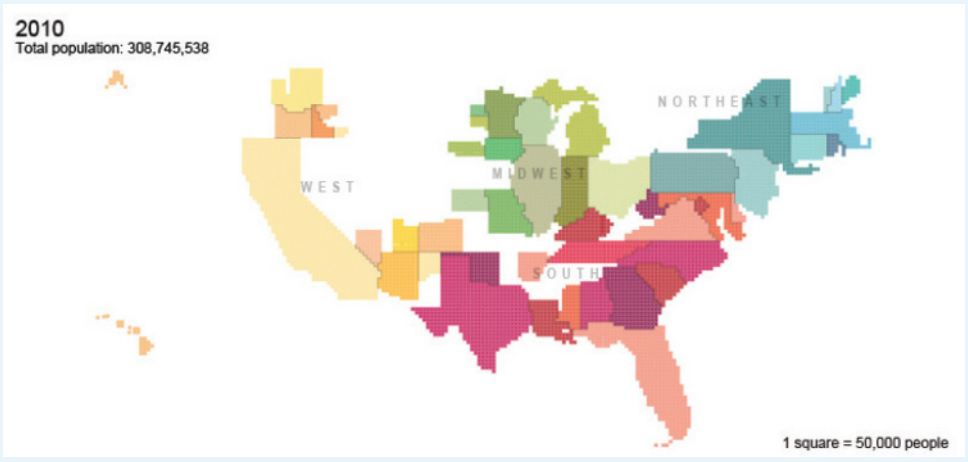


It may take Excel several minutes to populate the county-level choropleth map.

often leverages the audience's familiarity with the geography of the displayed regions to convey its message. In Figure 3.44, we observe the tiny size of many western states (Alaska, Nevada, Idaho, Nebraska, North Dakota, and South Dakota) and the white space paired with the audience's tacit knowledge of a standard U.S. map conveys the low population density in these areas.

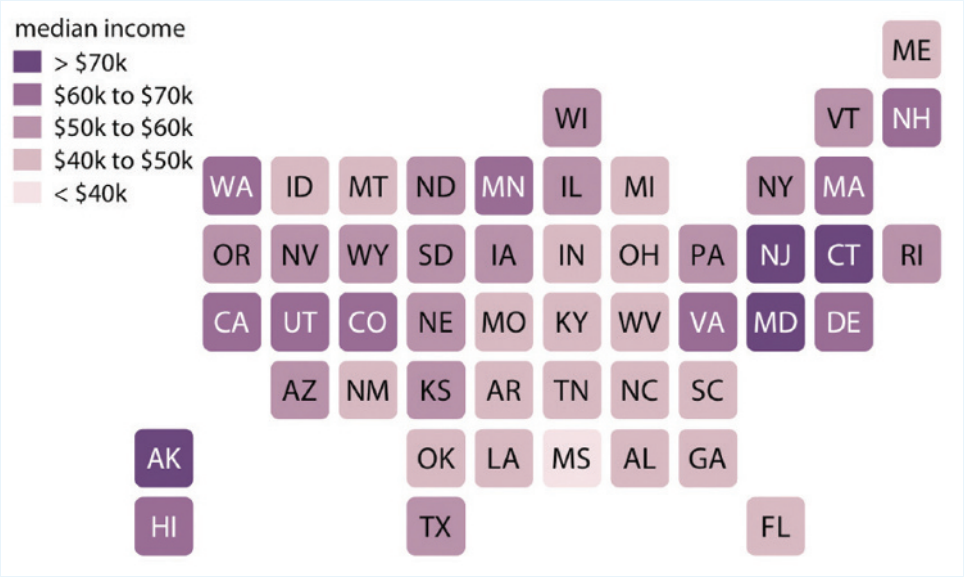
A strength of a cartogram is that the area displayed is proportional to the variable being measured, thus avoiding any misleading impressions. A weakness of a cartogram is that the sizing of the regions according to the displayed variable may distort enough to render

Figure 3.44 Cartogram of United States Displaying State Population



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.45 Equal-Area Cartogram of United States Displaying Median Income by State



the relative geographic positioning meaningless and the standard area-based geography unrecognizable. Figure 3.45 contains a specific type of cartogram, known as an **equal-area cartogram**. In comparison to the choropleth in Figure 3.42, the equal-area cartogram provides a balanced visual representation of each state while still maintaining fidelity to relative geographic positioning.

3.6 Data Dashboards

A **data dashboard** is a data-visualization tool that illustrates multiple metrics and automatically updates these metrics as new data become available. It is like an automobile's dashboard instrumentation that provides information on the vehicle's current speed, fuel level, and engine temperature so that a driver can assess current operating conditions and take effective action. Similarly, a data dashboard provides the important metrics that managers need to quickly assess the performance of their organization and react accordingly. In this section, we provide guidelines for creating effective data dashboards and an example application.

Principles of Effective Data Dashboards

Key performance indicators are sometimes referred to as key performance metrics (KPMs).

In an automobile dashboard, values such as current speed, fuel level, and oil pressure are displayed to give the driver a quick overview of current operating characteristics. In a business, the equivalent values are often indicative of the business's current operating characteristics, such as its financial position, the inventory on hand, customer service metrics, and the like. These values are typically known as **key performance indicators (KPIs)**. A data dashboard should provide timely summary information on KPIs that are important to the user, and it should do so in a manner that informs rather than overwhelms its user.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Ideally, a data dashboard should present all KPIs as a single screen that a user can quickly scan to understand the business's current state of operations. Rather than requiring the user to scroll vertically and horizontally to see the entire dashboard, it is better to create multiple dashboards so that each dashboard can be viewed on a single screen.

The KPIs displayed in the data dashboard should convey meaning to its user and be related to the decisions the user makes. For example, the data dashboard for a marketing manager may have KPIs related to current sales measures and sales by region, while the data dashboard for a Chief Financial Officer should provide information on the current financial standing of the company, including cash on hand, current debt obligations, and so on.

A data dashboard should call attention to unusual measures that may require attention, but not in an overwhelming way. Color should be used to call attention to specific values to differentiate categorical variables, but the use of color should be restrained. Too many different colors or colors that are too bright make the presentation distracting and difficult to read.

Applications of Data Dashboards

To illustrate the use of a data dashboard in decision making, we discuss an application involving the Grogan Oil Company which has offices located in three cities in Texas: Austin (its headquarters), Houston, and Dallas. Grogan's Information Technology (IT) call center, located in Austin, handles calls from employees regarding computer-related problems involving software, Internet, and e-mail issues. For example, if a Grogan employee in Dallas has a computer software problem, the employee can call the IT call center for assistance.

The data dashboard shown in Figure 3.46, developed to monitor the performance of the call center, combines several displays to track the call center's KPIs. The data presented are for the current shift, which started at 8:00 a.m. The stacked column chart in the upper left-hand corner shows the call volume for each type of problem (software, Internet, or email) over time. This chart shows that call volume is heavier during the first few hours of the shift, calls concerning email issues appear to decrease over time, and volume of calls regarding software issues are highest at midmorning.

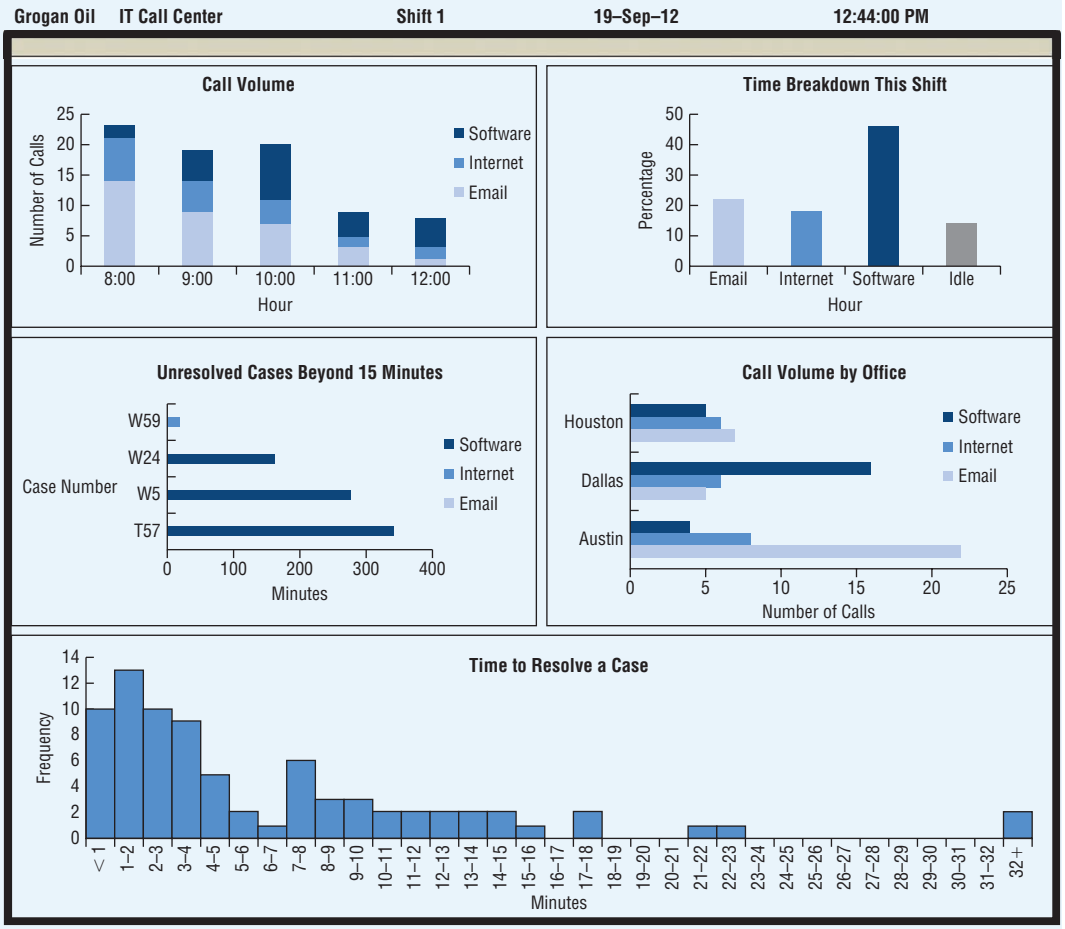
The column chart in the upper right-hand corner of the dashboard shows the percentage of time that call center employees spent on each type of problem or were idle (not working on a call). These top two charts are important displays in determining optimal staffing levels. For instance, knowing the call mix and how stressed the system is, as measured by percentage of idle time, can help the IT manager make sure that enough call center employees are available with the right level of expertise.

The clustered-bar chart in the middle right of the dashboard shows the call volume by type of problem for each of Grogan's offices. This allows the IT manager to quickly identify whether there is a particular type of problem by location. For example, the office in Austin seems to be reporting a relatively high number of issues with e-mail. If the source of the problem can be identified quickly, then the problem might be resolved quickly for many users all at once. Also, note that a relatively high number of software problems are coming from the Dallas office. In this case, the Dallas office is installing new software, resulting in more calls to the IT call center. Having been alerted to this by the Dallas office last week, the IT manager knew that calls coming from the Dallas office would spike, so the manager proactively increased staffing levels to handle the expected increase in calls.

For each unresolved case that was received more than 15 minutes ago, the bar chart shown in the middle left of the data dashboard displays the length of time for which each case has been unresolved. This chart enables Grogan to quickly monitor the key problem

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 3.46 Data Dashboard for the Grogan Oil Information Technology Call Center



DATA*file*
grogan

The construction of frequency distributions for quantitative and categorical data is discussed in Chapter 2.

cases and decide whether additional resources may be needed to resolve them. The worst case, T57, has been unresolved for over 300 minutes and is actually left over from the previous shift. Finally, the chart in the bottom panel shows the length of time required for resolved cases during the current shift. This chart is an example of a frequency distribution for quantitative data.

Throughout the dashboard, a consistent color coding scheme is used for problem type (e-mail, software, and internet). Other dashboard designs are certainly possible, and improvements could certainly be made to the design shown in Figure 3.46. However, what is important is that information is clearly communicated so that managers can improve their decision making.

The Grogan Oil data dashboard presents data at the operational level, is updated in real time, and is used for operational decisions such as staffing levels. Data dashboards may also be used at the tactical and strategic levels of management. For example, a sales manager could monitor sales by salesperson, by region, by product, and by customer. This would alert the sales manager to changes in sales patterns. At the highest level, a more strategic dashboard would allow upper management to quickly assess the financial health of the company by monitoring more aggregate financial, service-level, and capacity-utilization information.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notes + Comments

1. Interactive charts and tables can be created in Excel using PivotCharts and PivotTables, which can form the basis of simple data dashboards. The creation of more advanced data dashboards in Excel generally requires the use of macros written using Visual Basic for Applications (VBA). The use of VBA is beyond the scope of this textbook, but VBA is a powerful programming tool that can greatly increase the capabilities of Excel for analytics, including data visualization. Dedicated data visualization software packages, such as Tableau, Power BI, Domo, IBM Cognos, and R Shiny, make it much easier to create data dashboards.
2. The appendix to this chapter provides instructions for creating basic data visualizations in Tableau. Online appendices available for this text provide instructions for creating visualizations in other common analytics software.

Summary

In this chapter, we covered techniques and tools related to data visualization. We discussed several important principles related to creating effective data visualizations including preattentive attributes and data-ink ratio. We introduced techniques for enhancing visual presentations, such as decluttering tables and charts by removing unnecessary lines and presenting numerical values only to the precision necessary for analysis. We explained that tables can be preferable to charts for data visualization when the user needs to know exact numerical values. We introduced crosstabulation as a form of a table for two variables and explained how to use Excel to create a PivotTable.

We presented a variety of standard charts that can be used for many different applications of data visualization, including scatter charts, line charts (including sparklines), as well as bar and column charts. We explained that pie charts and three-dimensional charts are almost never preferred tools for data visualization and that bar (or column) charts are usually much more effective than pie charts. We introduced several visualizations that can be used to compare multiple variables including a bubble chart, a scatter chart matrix, and a table lens. We also discussed some specialized data-visualization charts, such as heat maps, treemaps, waterfall charts, stock charts, and parallel-coordinates plots. We presented choropleth maps and cartograms as visualizations that can be used with geospatial data, and we discussed the pros and cons of each of these types of visualization. We introduced data dashboards as a data-visualization tool that provides a summary of a firm's operations in visual form to allow managers to quickly assess the current operating conditions and to aid decision making.

Many other types of charts can be used for specific forms of data visualization, but we have covered many of the most-popular and most-useful ones. Data visualization is very important for helping analyze data and identify important relations and patterns. The effective design of tables and charts is also necessary to communicate data analysis to others. Tables and charts should be only as complicated as necessary to help the user understand the patterns and relationships in the data.

Glossary

Bar chart A graphical presentation that uses horizontal bars to display the magnitude of quantitative data. Each bar typically represents a class of a categorical variable.
Bubble chart A graphical presentation used to visualize three variables in a two-Dimensional graph. The two axes represent two variables, and the magnitude of the third variable is given by the size of the bubble.
Cartogram A map-like diagram that uses geographic positioning but purposefully represents map regions in a manner that does not necessarily correspond to land area.
Chart A visual method for displaying data; also called a graph or a figure.
Choropleth map A geographic visualization that uses shades of a color, different colors, or symbols to indicate the values of a variable associated with a region.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Clustered-column (or clustered-bar) chart A special type of column (bar) chart in which multiple bars are clustered in the same class to compare multiple variables; also known as a side-by-side-column (bar) chart.
Cognitive load The amount of effort necessary to accurately and efficiently process the information being communicated by a data visualization.
Column chart A graphical presentation that uses vertical bars to display the magnitude of quantitative data. Each bar typically represents a class of a categorical variable.
Crosstabulation A tabular summary of data for two variables. The classes of one variable are represented by the rows; the classes for the other variable are represented by the columns.
Data dashboard A data-visualization tool that updates in real time and gives multiple outputs.
Data-ink ratio The ratio of the amount of ink used in a table or chart that is necessary to convey information to the total amount of ink used in the table and chart. Ink used that is not necessary to convey information reduces the data-ink ratio.
Decluttering The act of removing the non-data-ink in the visualization that does not help the audience interpret the chart or table.
Equal-area cartogram A map-like diagram that uses relative geographic positioning of regions, but uses shapes of equal (or near-equal) size to represent regions.
Geospatial data Data that include information on the geographic location of each record.
Heat map A two-dimensional graphical presentation of data in which color shadings indicate magnitudes.
Hierarchical data Data that can be represented with a tree-like structure, where the branches of the tree lead to categories and subcategories.
High-low-close stock chart A graphical presentation that shows the high value, low value, and closing value of the price of a share of stock over time.
Key performance indicator (KPI) A metric that is crucial for understanding the current performance of an organization; also known as a key performance metric (KPM).
Line chart A graphical presentation of time series data in which the data points are connected by a line.
Parallel-coordinates plot A graphical presentation used to examine more than two variables in which each variable is represented by a different vertical axis. Each observation in a data set is plotted in a parallel-coordinates plot by drawing a line between the values of each variable for the observation.
Pie chart A graphical presentation used to compare categorical data. Because of difficulties in comparing relative areas on a pie chart, these charts are not recommended. Bar or column charts are generally superior to pie charts for comparing categorical data.
PivotChart A graphical presentation created in Excel that functions similarly to a PivotTable.
Preattentive attributes Features such as color, size and length that can be used in a data visualization to reduce the cognitive load required by the user to interpret the visualization.
PivotTable An interactive crosstabulation created in Excel.
Scatter chart A graphical presentation of the relationship between two quantitative variables. One variable is shown on the horizontal axis and the other on the vertical axis.
Scatter-chart matrix A graphical presentation that uses multiple scatter charts arranged as a matrix to illustrate the relationships among multiple variables.
Sparkline A special type of line chart that has no axes or labels used to indicate the trend of data but not magnitude.
Stacked-column chart A special type of column (bar) chart in which multiple variables appear on the same bar.
Stock chart A graphical presentation of stock prices over time.
Table lens A tabular-like visualization in which each column corresponds to a variable and the magnitude of a variable's values are represented by horizontal bars.
Treemap A graphical presentation that uses the size, color, and arrangement of rectangles to display the magnitudes of a quantitative variable for different categories, each of which are further decomposed into subcategories.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Trendline A line that provides an approximation of the relationship between variables in a chart.

Waterfall chart A graphical presentation that shows the cumulative effect of positive and negative changes on a variable of interest. The basis of the changes can be time or categories and changes are represented by columns anchored at the previous time or category's cumulative level.

Problems

1. **Sales Performance Bonuses.** A sales manager is trying to determine appropriate sales performance bonuses for the sales team this year. The following table contains the data relevant to determining the bonuses, but it is not easy to read and interpret. Reformat the table to improve readability and to help the sales manager make decisions about bonuses. **LO 1**

DATA*file* salesbonuses

Salesperson	Total Sales ($)	Average Performance Bonus Previous Years ($)	Customer Accounts	Years with Company
Smith, Michael	325000.78	12499.3452	124	14
Yu, Joe	13678.21	239.9434	9	7
Reeves, Bill	452359.19	21987.2462	175	21
Hamilton, Joshua	87423.91	7642.9011	28	3
Harper, Derek	87654.21	1250.1393	21	4
Quinn, Dorothy	234091.39	14567.9833	48	9
Graves, Lorrie	379401.94	27981.4432	121	12
Sun, Yi	31733.59	672.9111	7	1
Thompson, Nicole	127845.22	13322.9713	17	3

2. **Gross Domestic Product Values.** The following table shows an example of gross domestic product values for five countries over six years in equivalent U.S. dollars ($). **LO 1**

Gross Domestic Product (US $)						
Country	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Albania	7385937423	8105580293	9650128750	11592303225	10781921975	10569204154
Argentina	169725491092	198012474920	241037555661	301259040110	285070994754	339604450702
Australia	704453444387	758320889024	916931817944	982991358955	934168969952	1178776680167
Austria	272865358404	290682488352	336840690493	375777347214	344514388622	341440991770
Belgium	335571307765	355372712266	408482592257	451663134614	421433351959	416534140346

DATA*file* gdp_years

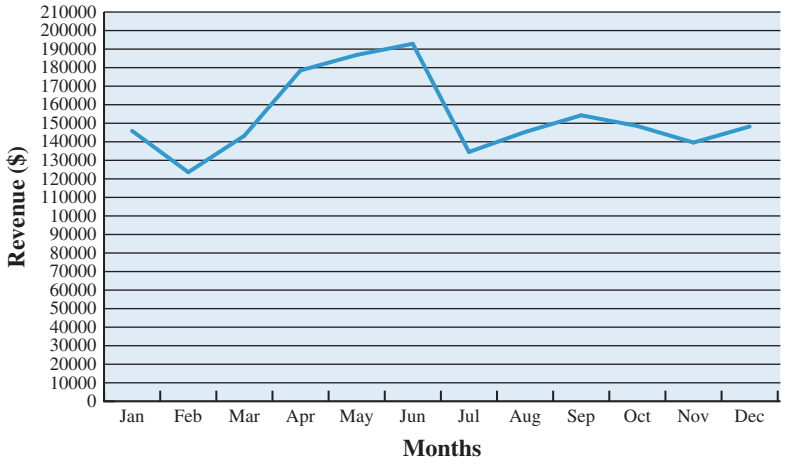
a. How could you improve the readability of this table?
b. The file *gdp_years* contains sample data from the United Nations Statistics Division on 30 countries and their GDP values from Year 1 to Year 6 in US $. Create a table that provides all these data for a user. Format the table to make it as easy to read as possible. (*Hint:* It is generally not important for the user to know GDP to an exact dollar figure. It is typical to present GDP values in millions or billions of dollars.)

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

3. **Monthly Revenue Data.** The following table provides monthly revenue values for Tedstar, Inc., a company that sells valves to large industrial firms. The monthly revenue data have been graphed using a line chart in the following figure. **LO 1, 4**

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Revenue ($)	145869	123576	143298	178505	186850	192850	134500	145286	154285	148523	139600	148235

DATA *file*
tedstar



a. What are the problems with the layout and display of this line chart?
b. Create a new line chart for the monthly revenue data at Tedstar, Inc. Format the chart to make it easy to read and interpret.

DATA *file*
majorsalary

4. **Business Graduates Salaries.** In the file *majorsalary*, data have been collected from 111 College of Business graduates on their monthly starting salaries. The graduates include students majoring in management, finance, accounting, information systems, and marketing. Create a PivotTable in Excel to display the number of graduates in each major and the average monthly starting salary for students in each major. **LO 2**
   a. Which major has the greatest number of graduates?
   b. Which major has the highest average starting monthly salary?
   c. Use the PivotTable to determine the major of the student with the highest overall starting monthly salary. What is the major of the student with the lowest overall starting monthly salary?
5. **Top U.S. Franchises.** *Entrepreneur* magazine ranks franchises. Among the factors that the magazine uses in its rankings are growth rate, number of locations, start-up costs, and financial stability. A recent ranking listed the top 20 U.S. franchises and the number of locations as follows:

Franchise	Number of U.S. Locations	Franchise	Number of U.S. Locations
Hampton Inns	1,864	Jan-Pro Franchising Intl. Inc.	12,394
ampm	3,183	Hardee's	1,901
McDonald's	32,805	Pizza Hut Inc.	13,281
7-Eleven Inc.	37,496	Kumon Math & Reading Centers	25,199
Supercuts	2,130	Dunkin' Donuts	9,947
Days Inn	1,877	KFC Corp.	16,224
Vanguard Cleaning Systems	2,155	Jazzercise Inc.	7,683
Servpro	1,572	Anytime Fitness	1,618
Subway	34,871	Matco Tools	1,431
Denny's Inc.	1,668	Stratus Building Solutions	5,018

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file* franchises

These data can be found in the file *franchises*. Create a PivotTable to summarize these data using classes 0–9,999, 10,000–19,999, 20,000–29,999, and 30,000–39,999 to answer the following questions. (*Hint:* Use Number of U.S. Locations as the Columns, and use Count of Number of U.S. Locations as the **Values** in the PivotTable.) **LO 2**

a. How many franchises have between 0 and 9,999 locations?
b. How many franchises have more than 30,000 locations?

DATA*file* mutualfunds

6. **Mutual Funds Data.** The file *mutualfunds* contains a data set with information for 45 mutual funds that are part of the *Morningstar Funds 500*. The data set includes the the five variables listed below. **LO 2**

Variable Name	Definition
Fund Type	Type of fund, labeled DE (Domestic Equity), IE (International Equity), and FI (Fixed Income)
Net Asset Value ($)	Closing price per share
Five-Year Average Return (%)	Average annual return for the fund over the past five years
Expense Ratio (%)	Percentage of assets deducted each fiscal year for fund expenses
Morningstar Rank	Risk adjusted star rating for each fund; Morningstar ranks go From a low of 1 Star to a high of 5 Stars

Note that Excel may display the column headings as 0–10, 10–20, 20–30, etc., but they should be interpreted as 0–9.99, 10–19.99, 20–29.99, etc.

a. Prepare a PivotTable that gives the frequency count of the data by Fund Type (Rows) and the five-year average annual return (Columns). Use classes of 0–9.99, 10–19.99, 20–29.99, 30–39.99, 40–49.99, and 50–59.99 for the Five-Year Average Return (%). (*Hint*: You should group the columns of the PivotTable by setting the **Starting at**: value to *0*, the **Ending at**: value to *59.99* and the **By**: value to *10*.)
b. What conclusions can you draw about the fund type and the average return over the past five years?

DATA*file* taxdata

7. **Tax Data by County.** The file *taxdata* contains information from federal tax returns filed in 2007 for all counties in the United States (3,142 counties in total). Create a PivotTable in Excel to answer the questions below. The PivotTable should have State Abbreviation as Row Labels. The Values in the PivotTable should be the sum of adjusted gross income for each state. **LO 2**

a. Sort the PivotTable data to display the states with the smallest sum of adjusted gross income on top and the largest on the bottom. Which state had the smallest sum of adjusted gross income? What is the total adjusted gross income for federal tax returns filed in this state with the smallest total adjusted gross income? (*Hint:* To sort data in a PivotTable in Excel, right-click any cell in the PivotTable that contains the data you want to sort, and select **Sort**.)
b. Add the County Name to the Row Labels in the PivotTable. Sort the County Names by Sum of Adjusted Gross Income with the lowest values on the top and the highest values on the bottom. Filter the Row Labels so that only the state of Texas is displayed. Which county had the smallest sum of adjusted gross income in the state of Texas? Which county had the largest sum of adjusted gross income in the state of Texas?
c. Click on **Sum of Adjusted Gross Income** in the **Values** area of the PivotTable in Excel. Click **Value Field Settings…**. Click the tab for **Show Values As**. In the **Show values as** box, select **% of Parent Row Total**. Click **OK**. This displays the adjusted gross income reported by each county as a percentage of the total state adjusted gross income. Which county has the highest percentage adjusted gross income in the state of Texas? What is this percentage?
d. Remove the filter on the Row Labels to display data for all states. What percentage of total adjusted gross income in the United States was provided by the state of New York?

DATA*file* bankfailures

8. **Federally Insured Bank Failures.** The file *bankfailures* contains data on failures of federally insured banks between 2000 and 2012. Create a PivotTable in Excel to answer the following questions. The PivotTable should group the closing dates of the banks into yearly bins and display the counts of bank closures each year in columns

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

of Excel. Row labels should include the bank locations and allow for grouping the locations into states or viewing by city. You should also sort the PivotTable so that the states with the greatest number of total bank failures between 2000 and 2012 appear at the top of the PivotTable. **LO 2**

a. Which state had the greatest number of federally insured bank closings between 2000 and 2012?
b. How many bank closings occurred in the state of Nevada (NV) in 2010? In what cities did these bank closings occur?
c. Use the PivotTable's filter capability to view only bank closings in California (CA), Florida (FL), Texas (TX), and New York (NY) for the years 2009 through 2012. What is the total number of bank closings in these states between 2009 and 2012?
d. Using the filtered PivotTable from part c, what city in Florida had the greatest number of bank closings between 2009 and 2012? How many bank closings occurred in this city?
e. Create a PivotChart to display a column chart that shows the total number of bank closings in each year from 2000 through 2012 in the state of Florida. Adjust the formatting of this column chart so that it best conveys the data. What does this column chart suggest about bank closings between 2000 and 2012 in Florida? Discuss. (*Hint:* You may have to switch the row and column labels in the PivotChart to get the best presentation for your PivotChart.)

9. **Scatter Chart and Trendline.** The following 20 observations are for two quantitative variables, x and y. **LO 3, 5**

DATA*file* scatter

Observation	x	y	Observation	x	y
1	−22	22	11	−37	48
2	−33	49	12	34	−29
3	2	8	13	9	−18
4	29	−16	14	−33	31
5	−13	10	15	20	−16
6	21	−28	16	−3	14
7	−13	27	17	−15	18
8	−23	35	18	12	17
9	14	−5	19	−20	−11
10	3	−3	20	−7	−22

a. Create a scatter chart for these 20 observations.
b. Fit a linear trendline to the 20 observations. What can you say about the relationship between the two quantitative variables?

DATA*file* fortune500

10. **Profits and Market Capitalizations.** The file *fortune500* contains data for profits and market capitalizations from a recent sample of firms in the Fortune 500. **LO 3, 5**
a. Prepare a scatter diagram to show the relationship between the variables Market Capitalization and Profit in which Market Capitalization is on the vertical axis and Profit is on the horizontal axis. Comment on any relationship between the variables.
b. Create a trendline for the relationship between Market Capitalization and Profit. What does the trendline indicate about this relationship?

DATA*file* colleges

11. **Exploring Private Colleges**. The file *colleges* contains the following data for the sample of 102 private colleges: year founded, tuition and fees (not including room and board), and the percentage of undergraduates who obtained their degree within six years (source: *The World Almanac*). **LO 3, 5**
a. Create a scatter chart to explore the relationship between tuition and percent who graduate.
b. Create a linear trendline and comment on any apparent relationship.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

12. **Vehicle Production Data.** The International Organization of Motor Vehicle Manufacturers (officially known as the Organisation Internationale des Constructeurs d'Automobiles, OICA) provides data on worldwide vehicle production by manufacturer. The following table shows vehicle production numbers for four different manufacturers for five recent years. Data are in millions of vehicles. **LO 4**

DATA*file* autoproduction

	Production (millions of vehicles)				
Manufacturer	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**
Toyota	8.04	8.53	9.24	7.23	8.56
GM	8.97	9.35	8.28	6.46	8.48
Volkswagen	5.68	6.27	6.44	6.07	7.34
Hyundai	2.51	2.62	2.78	4.65	5.76

a. Construct a line chart for the time series data for years 1 through 5 showing the number of vehicles manufactured by each automotive company. Show the time series for all four manufacturers on the same graph.
b. What does the line chart indicate about vehicle production amounts from years 1 through 5? Discuss.
c. Construct a clustered-bar chart showing vehicles produced by automobile manufacturer using the year 1 through 5 data. Represent the years of production along the horizontal axis, and cluster the production amounts for the four manufacturers in each year. Which company is the leading manufacturer in each year?

13. **Comparing Package Delivery Drivers**. Red Sky Delivery performs "last-mile" delivery services for online retailers such as Amazon. Red Sky employs delivery drivers who deliver packages to individual residence and business locations from a local transportation hub. Red Sky measures several delivery driver performance metrics, including number of delivery stops completed per eight-hour shift. The table below provides data on nine Red Sky delivery drivers and their average number of packages delivered per shift over the previous 30 days. **LO 4**

DATA*file* redsky

Delivery Driver	**Average Number of Delivery Stops Completed (per shift)**
Amy Urbaczewski	92.87
Sally Melouk	110.78
Brenda Barnes	114.20
Jonathan Payne	132.50
Bruce Wheeler	148.20
Cam Madsen	87.51
Sheila Stevens	139.82
Grant Inhara	154.23
Finn Helton	109.11

a. Using the data in the file *redsky*, create a column chart to display the information in the table above. Add data labels that show the average number of delivery stops completed for each driver.
b. Create a sorted column chart to make it easier to see which drivers have the highest and lowest average number of delivery stops per shift.

14. **Beverage Distributor Revenue.** The file *beveragedistributor* contains data on the monthly revenue for a large regional distributor of beverages. (*Hint*: You can use a PivotChart to easily aggregate the sales amounts to different levels.) **LO 2, 4**

DATA*file* beveragedistributor

a. Create a line chart to depict the revenue time series at the annual level. What insight on beverage sales does this chart reveal?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

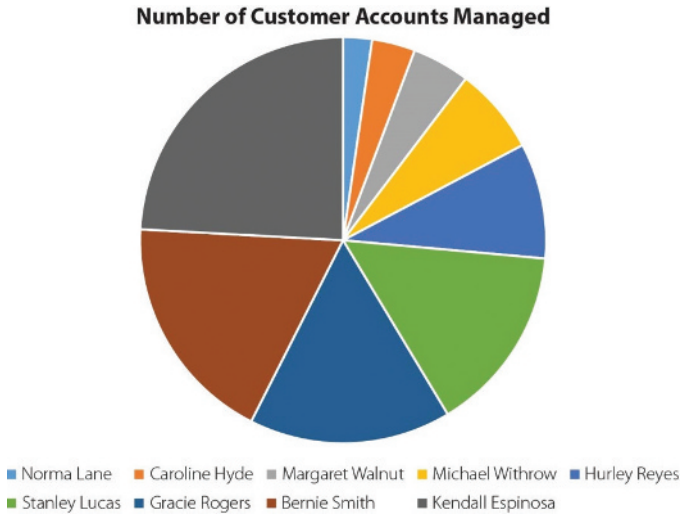
b. Create a line chart to depict the revenue time series at the quarterly level. What insight on beverage sales does this chart reveal?
c. Create a line chart to depict the revenue time series at the monthly level. What insight on beverage sales does this chart reveal?
d. Comparing the line charts created in parts (a), (b), and (c) discuss the importance of choosing the level of aggregation for time-series data?

15. **Term Life Insurance.** The following table contains sales totals for the top six term life insurance salespeople at American Insurance. LO 4, 6

Salesperson	Contracts Sold
Harish	24
David	41
Kristina	19
Steven	23
Tim	53
Mona	39

a. Create a column chart to display the information in the table above. Format the column chart to best display the data by adding axes labels, a chart title, etc.
b. Sort the values in Excel so that the column chart is ordered from most contracts sold to fewest.
c. Insert data labels to display the number of contracts sold for each salesperson above the columns in the column chart created in part a.

16. **Comparing Managing Consultants**. Blatt Consulting Services is interested in comparing the number of customer accounts managed by its consultants. The chart below is a default chart produced in Excel. LO 1, 4



DATA*file*
blattconsulting

a. Which preattentive attributes are used in the pie chart?
b. What preattentive attributes would you use in redesigning this chart to improve its clarity for the audience?
c. Use the data in the file *blattconsulting* to create a new data visualization that uses a different type of chart and improved use of preattentive attributes to allow for easier comparison of the number of accounts managed by each consultant.

17. **Engine Type Preference.** An automotive company is considering the introduction of a new model of sports car that will be available in four-cylinder and six-cylinder engine types. A sample of customers who were interested in this new model were asked to indicate their preference for an engine type for the new model of automobile. The customers were also asked to indicate their preference for exterior color from four choices: red,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

black, green, and white. Consider the data in the table shown below regarding the customer responses to answer the following questions. **LO 6**

DATA*file* new_auto

	Four Cylinders	Six Cylinders
Red	143	857
Black	200	800
Green	321	679
White	420	580

a. Construct a clustered-column chart with exterior color as the horizontal variable.
b. What can we infer from the clustered-column chart in part (a)?

18. **Smartphone Ownership.** Consider the survey results in the following table regarding smartphone ownership by age to answer the following questions. **LO 6**

DATA*file* smartphone

Age Category	Smartphone (%)	Other Cell Phone (%)	No Cell Phone (%)
18–24	49	46	5
25–34	58	35	7
35–44	44	45	11
45–54	28	58	14
55–64	22	59	19
65+	11	45	44

a. Construct a stacked-column chart to display the survey data on type of cell-phone ownership. Use Age Category as the variable on the horizontal axis.
b. Construct a clustered-column chart to display the survey data. Use Age Category as the variable on the horizontal axis.
c. What can you infer about the relationship between age and smartphone ownership from the column charts in parts (a) and (b)? Which column chart (stacked or clustered) is best for interpreting this relationship? Why?

19. **Store Manager Tasks.** The Northwest regional manager of Logan Outdoor Equipment Company has conducted a study to determine how store managers are allocating their time. A study was undertaken over three weeks that collected the following data related to the percentage of time each store manager spent on the tasks of attending required meetings, preparing business reports, customer interaction, and being idle. The results of the data collection appear in the following table. **LO 4, 6**

DATA*file* logan

		Attending Required Meetings (%)	Tasks Preparing Business Reports (%)	Customer Interaction (%)	Idle (%)
Locations	Seattle	32	17	37	14
	Portland	52	11	24	13
	Bend	18	11	52	19
	Missoula	21	6	43	30
	Boise	12	14	64	10
	Olympia	17	12	54	17

a. Create a stacked-bar chart with locations along the vertical axis. Reformat the bar chart to best display these data by adding axis labels, a chart title, and so on.
b. Create a clustered-bar chart with locations along the vertical axis and clusters of tasks. Reformat the bar chart to best display these data by adding axis labels, a chart title, and the like.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Create multiple bar charts in which each location becomes a single bar chart showing the percentage of time spent on tasks. Reformat the bar charts to best display these data by adding axis labels, a chart title, and so forth.
d. Which visualization (stacked-bar chart, clustered-bar chart, or multiple bar charts) is preferred for these data? Why?
e. What can we infer about the differences among how store managers are allocating their time at the different locations?

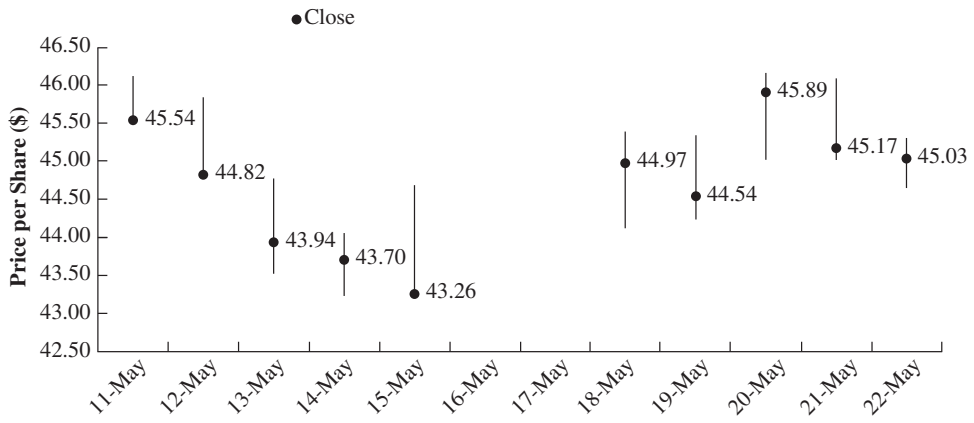
20. **R&D Project Portfolio.** The Ajax Company uses a portfolio approach to manage their research and development (R&D) projects. Ajax wants to keep a mix of projects to balance the expected return and risk profiles of their R&D activities. Consider a situation in which Ajax has six R&D projects as characterized in the table. Each project is given an expected rate of return and a risk assessment, which is a value between 1 and 10, where 1 is the least risky and 10 is the most risky. Ajax would like to visualize their current R&D projects to keep track of the overall risk and return of their R&D portfolio. **LO 7**

DATA*file* ajax

Project	Expected Rate of Return (%)	Risk Estimate	Capital Invested ($ millions)
1	12.6	6.8	6.4
2	14.8	6.2	45.8
3	9.2	4.2	9.2
4	6.1	6.2	17.2
5	21.4	8.2	34.2
6	7.5	3.2	14.8

a. Create a bubble chart in which the expected rate of return is along the horizontal axis, the risk estimate is on the vertical axis, and the size of the bubbles represents the amount of capital invested. Format this chart for best presentation by adding axis labels and labeling each bubble with the project number.
b. The efficient frontier of R&D projects represents the set of projects that have the highest expected rate of return for a given level of risk. In other words, any project that has a smaller expected rate of return for an equivalent, or higher, risk estimate cannot be on the efficient frontier. From the bubble chart in part a, which projects appear to be located on the efficient frontier?

21. **Coca-Cola Stock Prices.** The following stock chart shows stock price performance for Coca-Cola over a two-week period. Note that May 16 and May 17 are a Saturday and a Sunday and are non-trading days. **LO 8**



a. Which day seems to have the most intra-day price variability?
b. What is the closing price on May 22?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. If you bought 100 shares at the closing price on May 19 and sold all of those shares at the closing price on May 22, how much did you gain or lose (ignoring any transaction fees)?

DATA*file* bravman

22. **Customer Shopping Data.** Bravman Clothing sells high-end clothing products and is launching a service in which they use a same-day courier service to deliver purchases that customers have made by working with one of their personal stylists on the phone. In a pilot study, Bravman collected 25 observations consisting of the wait time the customer experienced during the order process, the customer's purchase amount, the customer's age, and the customer's credit score. The file *bravman* contains the data for these 25 observations. **LO 7**
    a. Construct a table lens on these data sorted in decreasing order with respect to the values of the Purchase Amount variable.
    b. What can you infer about the relationship between Purchase Amount and the other variables based on the table lens?

DATA*file* bravman

23. **Customer Shopping Data (Revisited).** In this problem we revisit the data from Bravman Clothing in Problem 22. The file *bravm*an which contains 25 observations consisting of the wait time the customer experienced during the order process, the customer's purchase amount, the customer's age, and the customer's credit score. **LO 7**
    a. Create a scatter-chart matrix for these data. Include the variables wait time, purchase amount, customer age, and credit score. (*Note*: Excel does not contain an automatic tool to generate a scatter chart matrix. This problem can be solved by either using a different software package or by creating the scatter charts individually in Excel.)
    b. What can you infer about the relationships between each pair of variables based on the scatter chart matrix?

DATA*file* surveyresults

24. **Marketing Survey Results.** Heat maps can be very useful for identifying missing data values in moderate to large data sets. The file *surveyresults* contains the responses from a marketing survey: 108 individuals responded to the survey of 10 questions. Respondents provided answers of 1, 2, 3, 4, or 5 to each question, corresponding to the overall satisfaction on 10 different dimensions of quality. However, not all respondents answered every question. **LO 8**
    a. To find the missing data values, create a heat map in Excel that shades the empty cells a different color. Use Excel's Conditional Formatting function to create this heat map. (*Hint:* Click on **Conditional Formatting** in the **Styles** group in the **Home** tab. Select **Highlight Cells Rules** and click **More Rules…**. Then enter **Blanks** in the **Format only cells with:** box. Select a format for these blank cells that will make them obviously stand out.)
    b. For each question, which respondents did not provide answers? Which question has the highest nonresponse rate?

25. **Revenues of Web Development Companies.** The following table shows monthly revenue for six different web development companies. **LO 8**

DATA*file* webdevelop

	Revenue ($)					
Company	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
Blue Sky Media	8,995	9,285	11,555	9,530	11,230	13,600
Innovate Technologies	18,250	16,870	19,580	17,260	18,290	16,250
Timmler Company	8,480	7,650	7,023	6,540	5,700	4,930
Accelerate, Inc.	28,325	27,580	23,450	22,500	20,800	19,800
Allen and Davis, LLC	4,580	6,420	6,780	7,520	8,370	10,100
Smith Ventures	17,500	16,850	20,185	18,950	17,520	18,580

    a. Use Excel to create sparklines for sales at each company.
    b. Which companies have generally decreasing revenues over the six months? Which company has exhibited the most consistent growth over the six months? Which companies have revenues that are both increasing and decreasing over the six months?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Use Excel to create a heat map for the revenue of the six companies. Do you find the heat map or the sparklines to be better at communicating the trend of revenues over the six months for each company? Why?

DATA*file* netmigration

26. **Regional Gains and Losses in Population.** The United States Census Bureau tracks shifts in population by each state and region in the United States. The net migration rate is an indicator of the movement of people from one area to another. A positive net migration rate means more people moved to the area than moved away, and a negative net migration rate means more people left the area than moved to the area. Mathematically, the net migration rate N is defined as follows.

$$N=\frac{(I-E)}{M}\times 1000$$

Where
I = number of people moving to the area in the year under consideration
E = the number of people moving away from the area in the year under consideration
M = the mid-year population of the area

The file *netmigration* contains net migration rates for four regions of the United States. **LO 8**

a. Create a heat map using conditional formatting with the **Blue-White-Red Color Scale**.
b. Which regions are losing population? Which regions are gaining population?

DATA*file* bellevuebakery

27. **Income Statement.** An income statement is a summary of a company's revenues and costs over a given period time. The data in the file *bellevuebakery* is an example of an income statement. It contains the revenues and costs for last year for Bellevue Bakery. Revenues include gross sales and other income. Costs include returns, cost of goods sold, advertising, salaries/wages, other operating expenses, and taxes. In the income statement there are intermediate calculations defined as follows. **LO 8**

Net Sales = Gross Sales − Returns
Gross Profit = Net Sales − Other Income − Cost of Goods Sold
Net Income Before Taxes = Gross Profit − Advertising − Salaries/Wages − Other Operating Expenses
Net Income = Net Income Before Taxes − Taxes

a. Create a waterfall chart of the income statement for Bellevue Bakery. In this waterfall chart, for each intermediate calculation "set as total."
b. Create another waterfall chart of the income statement for Bellevue Bakery, but do not include any of the entries corresponding to intermediate calculations except Net Income which you should "set as total."

28. **NFL Attendance.** Below is a sample of the data in the file *nfl_attendance* which contains the 32 teams in the National Football League, their conference affiliation, their division, and their average home attendance. **LO 1, 6, 8**

DATA*file* nfl_attendance

Conference	Division	Team	Average Home Attendance
AFC	West	Oakland	54,584
AFC	West	Los Angeles Chargers	57,024
NFC	North	Chicago	60,368
AFC	North	Cincinnati	60,511
NFC	South	Tampa Bay	60,624
NFC	North	Detroit	60,792
AFC	South	Jacksonville	61,915

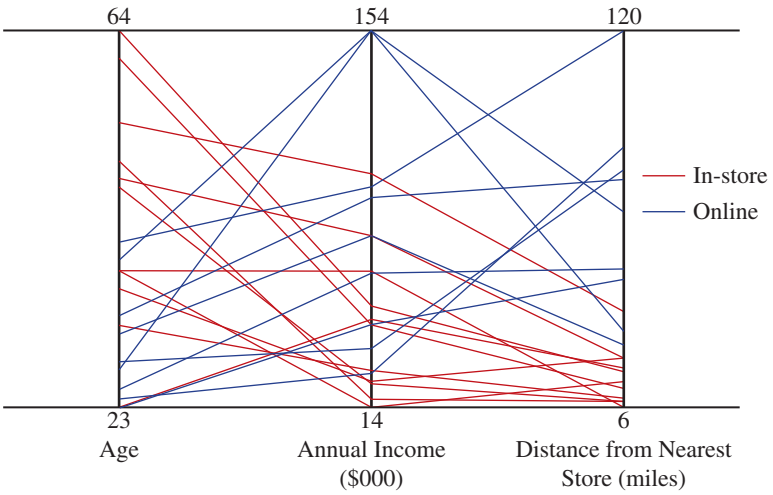
a. Create a treemap using these data that separates the teams into their conference affiliations (NFC and AFC) and uses size to represent each team's average home attendance. Note that you will need to sort the data in Excel by Conference to properly create a treemap.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Create a sorted bar chart that compares the average home attendance for each team.
c. Comment on the advantages and disadvantages of each type of chart for these data. Which chart best displays these data and why?

DATA*file* management25

29. **Top Management.** The Drucker Institute ranks corporations for managerial effectiveness based on a composite score derived from the following five factors: customer satisfaction, employee engagement and development, innovation, financial strength, and social responsibility. The file *management25* contains the top 25 companies in the Institute's ranking based on the composite score (source: *The Wall Street Journal*). For each company, the industry sector, the company name, and the composite score are given. **LO 8**
    a. Create a treemap chart using these data with the sector being the category, company being the subcategory, and the composite score being the quantitative variable. Use "Management Top 25" for the chart title. (*Hint*: Be sure to first sort the data by sector.)
    b. Which sector has the most companies in the Top 25? Which company has the highest composite score in this sector?
30. **Online Customers versus In-Store Customers.** Zeitler's Department Stores sells its products online and through traditional brick-and-mortar stores. The following parallel-coordinates plot displays data from a sample of 20 customers who purchased clothing from Zeitler's either online or in-store. The data include variables for the customer's age, annual income, and the distance from the customer's home to the nearest Zeitler's store. According to the parallel-coordinates plot, how are online customers differentiated from in-store customers? **LO 8**



This problem cannot be completed with native Excel and requires a software package (e.g., Tableau, R) capable of constructing a parallel-coordinates plot.

DATA*file* zeitlers_electronics

31. **Customers Who Purchase Electronic Equipment.** The file *zeitlers_electronics* contains data on customers who purchased electronic equipment either online or in-store from Zeitler's Department Stores. **LO 8**
    a. Create a parallel-coordinates plot for these data. Include vertical axes for the customer's age, annual income, and distance from nearest store. Color the lines by the type of purchase made by the customer (online or in-store).
    b. How does this parallel-coordinates plot compare to the one shown in Problem 23 for clothing purchases? Does the division between online and in-store purchasing habits for customers buying electronics equipment appear to be the same as for customers buying clothing?
    c. Parallel-coordinates plots are very useful for interacting with your data to perform analysis. Filter the parallel-coordinates plot so that only customers whose homes are more than 40 miles from the nearest store are displayed. What do you learn from the parallel-coordinates plot about these customers?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

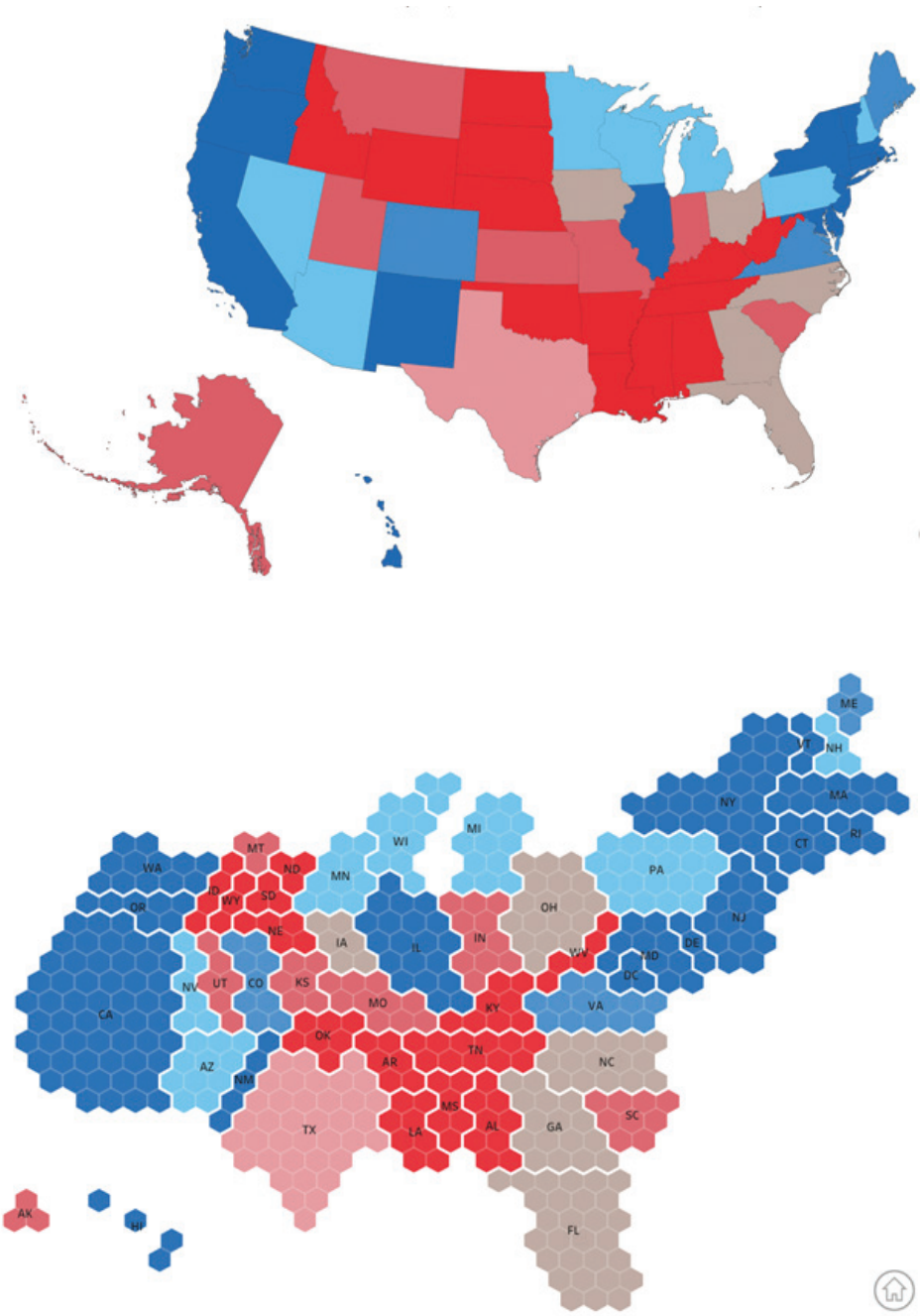
DATA*file*
species

32. **Biodiversity Preservation.** Ecologists often measure the biodiversity of a region by the number of distinct species that exist in the region. Nature reserves are lands specifically designated by the government to help maintain biodiversity. Care must be taken when setting up a network of nature reserves so that the maximum number of species can exist in the network. Geography matters as well, as putting reserves too close together might subject the entire network to risks, such as devastation from wildfires. The initial step in this type of planning usually involves mapping the number of species that exist in each region. The file *species* contains an estimate of the number of unique species that exist in each of the 50 states in the United States. **LO 9**
    a. Create a choropleth map that displays number of species by state. Add data labels to the choropleth map to show the number of species in each state.
    b. Comment on the distribution of species over the United States. Which regions of the United States have relatively many species? Which regions have relatively few species?
    c. Which two states have the most species? Why might this be expected?
    d. The choropleth map in part (a) helps to identify which states have the most species. However, how could one construct a choropleth map that identifies the states with the highest density of species?

DATA*file*
nebraskabeef

33. **Omaha Steaks**. Omaha Steaks is a food retailing company that specializes in mail order delivery of beef. The company is headquartered in Omaha Nebraska and was founded as a butcher shop in 1917. Omaha Steaks wants to create a visualization displaying the state's geographic distribution of beef cattle. Data on the number of beef cattle in each county in Nebraska and the size of each county (in square miles) are contained in the file *nebraskabeef*. **LO 9**
    a. Create a choropleth map that shades each county in Nebraska relative to the density of beef cattle (as measured by head per square mile).
    b. Why might it be preferable to shade each county based on density of beef cattle (as measured by head per square mile) rather than just number of beef cattle?
34. **Electoral Vote Maps**. The President of the United States is elected based on the electoral college. Each state in the United States has a specified number of electoral votes. Typically, the candidate that wins the popular vote in a state receives all the electoral votes assigned to that state (a few states like Maine and Nebraska split their electoral votes among the candidates based on the results of the popular vote). Suppose that a polling firm would like to create a visual for an upcoming Presidential election to show the relative standing of the Democratic and Republican candidates based on the most recent projections for the upcoming election. The first visualization shown below represents a choropleth map based on the latest projections. States that are shown in light (dark) blue are considered slightly (solidly) Democratic states. States that are shown in light (dark) red are considered slightly (solidly) Republican states. States that are shown in tan are considered to be tossup states that could be won by either the Democratic or the Republican candidate. The second visualization represents a cartogram where each grid shown represents one electoral vote. Electoral votes are based on population, not geographic size. Thus, large states (in area) with small populations, like Montana (MT), are large in the choropleth map but small in the cartogram (because they have relatively few electoral votes). Compare the choropleth map to the cartogram. What advantages are provided by the cartogram compared to the choropleth map? **LO 9**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



35. **Radiological Imaging Services Clinics.** Aurora Radiological Services is a health care clinic that provides radiological imaging services (such as MRIs, X-rays, and CAT scans) to patients. It is part of Front Range Medical Systems that operates clinics throughout the state of Colorado. **LO 10**
    a. What type of key performance indicators and other information would be appropriate to display on a data dashboard to assist the Aurora clinic's manager in making daily staffing decisions for the clinic?
    b. What type of key performance indicators and other information would be appropriate to display on a data dashboard for the CEO of Front Range Medical Systems who oversees the operation of multiple radiological imaging clinics?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Case Problem 1: Pelican Stores

Pelican Stores, a division of National Clothing, is a chain of clothing stores operating throughout the country. The chain recently ran a promotion in which discount coupons were sent to customers of other National Clothing stores. Data collected for a sample of 100 in-store credit card transactions at Pelican Stores during one day while the promotion was running are contained in the file *pelicanstores*. Table 3.17 shows a portion of the data set. The Proprietary Card method of payment refers to charges made using a National Clothing charge card. Customers who made a purchase using a discount coupon are referred to as promotional customers and customers who made a purchase but did not use a discount coupon are referred to as regular customers. Because the promotional coupons were not sent to regular Pelican Stores customers, management considers the sales made to people presenting the promotional coupons as sales it would not otherwise make. Of course, Pelican also hopes that the promotional customers will continue to shop at its stores.

Most of the variables shown in Table 3.17 are self-explanatory, but two of the variables require some clarification.

Items	The total number of items purchased
Net Sales	The total amount ($) charged to the credit card

Pelican's management would like to use this sample data to learn about its customer base and to evaluate the promotion involving discount coupons. **LO 2, 3, 4, 5, 6**

DATA*file*
pelicanstores

Table 3.17 Data for a Sample of 100 Credit Card Purchases at Pelican Stores

Customer	Type of Customer	Items	Net Sales	Method of Payment	College Graduate	Marital Status	Age
1	Regular	1	39.50	Discover	No	Married	32
2	Promotional	1	102.40	Proprietary Card	Yes	Married	36
3	Regular	1	22.50	Proprietary Card	Yes	Married	32
4	Promotional	5	100.40	Proprietary Card	Yes	Married	28
5	Regular	2	54.00	MasterCard	Yes	Married	34
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
96	Regular	1	39.50	MasterCard	Yes	Married	44
97	Promotional	9	253.00	Proprietary Card	Yes	Married	30
98	Promotional	10	287.59	Proprietary Card	Yes	Married	52
99	Promotional	2	47.60	Proprietary Card	Yes	Married	30
100	Promotional	1	28.44	Proprietary Card	Yes	Married	44

Managerial Report

Use the tabular and graphical methods of descriptive statistics to help management develop a customer profile and to evaluate the promotional campaign. At a minimum, your report should include the following:

1. Percent frequency distributions for each of the key variables: number of items purchased, net sales, method of payment, college graduate, marital status, and age.
2. A sorted bar chart showing the number of customer purchases attributable to the method of payment.
3. A crosstabulation of type of customer (regular or promotional) versus net sales. Comment on any similarities of differences observed.

Percent frequency distributions were introduced in Section 2.4. You can use Excel to create percent frequency distributions using PivotTables.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4. A scatter chart to explore the relationship between net sales and customer age.
5. A chart to examine whether the relationship between net sales and age depends on the marital status of the customer.
6. A side-by-side bar chart to examine the method of payment by customer type (regular or promotional). Comment on any differences you observe between the methods of payments used by the different types of customers.

Case Problem 2: Movie Theater Releases

The movie industry is a competitive business. More than 50 studios produce hundreds of new movies for theater release each year, and the financial success of each movie varies considerably. The opening weekend gross sales ($ millions), the total gross sales ($ millions), the number of theaters the movie was shown in, and the number of weeks the movie was in release are common variables used to measure the success of a movie released to theaters. Data collected for the top 100 theater movies released in 2016 are contained in the file *movies2016* (Box Office Mojo website). Table 3.18 shows the data for the first 10 movies in this file. **LO 2, 3**

Managerial Report

Use the tabular and graphical methods of descriptive statistics to learn how these variables contribute to the success of a motion picture. Include the following in your report.

1. Tabular and graphical summaries for each of the four variables along with a discussion of what each summary tells us about the movies that are released to theaters.
2. A scatter diagram to explore the relationship between Total Gross Sales and Opening Weekend Gross Sales. Discuss.
3. A scatter diagram to explore the relationship between Total Gross Sales and Number of Theaters. Discuss.
4. A scatter diagram to explore the relationship between Total Gross Sales and Number of Weeks in Release. Discuss.

DATA*file*
movies2016

Table 3.18 Performance Data for Ten 2016 Movies Released to Theaters

Movie Title	Opening Gross Sales ($ millions)	Total Gross Sales ($ millions)	Number of Theaters	Weeks in Release
Rogue One: A Star Wars Story	155.08	532.18	4,157	20
Finding Dory	135.06	486.30	4,305	25
Captain America: Civil War	179.14	408.08	4,226	20
The Secret Life of Pets	104.35	368.38	4,381	25
The Jungle Book	103.26	364.00	4,144	24
Deadpool	132.43	363.07	3,856	18
Zootopia	75.06	341.27	3,959	22
Batman v Superman: Dawn of Justice	166.01	330.36	4,256	12
Suicide Squad	133.6	325.10	4,255	14
Sing	35.26	270.40	4,029	20

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 4

Data Wrangling: Data Management and Data Cleaning Strategies

Contents

Analytics in Action:
Wrangling Data at Procter & Gamble
4.1 Discovery
Accessing Data
The Format of the Raw Data
4.2 Structuring
Data Formatting
Arrangement of Data
Splitting a Single Field into Multiple Fields
Combining Multiple Fields into a Single Field
4.3 Cleaning
Missing Data
Identification of Erroneous Outliers, Other Erroneous Values, and Duplicate Records
4.4 Enriching
Subsetting Data
Supplementing Data
Enhancing Data
4.5 Validating and Publishing
Validating
Publishing

Summary 188
Glossary 189
Problems 190

Available in Cengage eBook:
Appendix: Importing Delimited Files into R
Appendix: Working with Records in R
Appendix: Working with Fields in R
Appendix: Unstacking and Stacking Data in R

Learning Objectives

After completing this chapter, you will be able to:

LO 1 List and explain the steps of data wrangling.

LO 2 Import raw data from csv files, space delimited files, and tab delimited files.

LO 3 Reformat stacked data to unstacked data and reformat unstacked data to stacked data.

LO 4 Split a single field into multiple fields and combine multiple fields into a single field.

LO 5 Find and count missing values, find extreme values outside of a reasonable range, and find duplicate records.

LO 6 Subset data to isolate records that have values of interest for specific fields.

LO 7 Add fields and observations to raw data.

LO 8 Combine data for similar records from multiple files.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

Wrangling Data at Procter & Gamble*

Procter & Gamble (P&G) is one of the largest consumer packaged goods companies in the world. P&G serves consumers around the world with a portfolio of trusted, quality, leading brands including Charmin, Crest, Gillette, Olay, Oral-B, Pantene, and Tide. To operate at the speed and efficiency required by today's market, P&G must process large volumes of new data daily. Before any data set is used to produce business insights at P&G, it goes through a complex data governance and data wrangling process. After legal and information security aspects of data acquisition and data usage are clarified, a dedicated team takes responsibility for the data and enters relevant information into P&G's data catalog. Before data are used for downstream applications, the data are validated and loaded into the data model, which meets company standards and is easier to understand by subscribers of the data.

In order to meet the unique needs of each downstream analysis, the data are filtered and aggregated to ensure that only relevant data are processed in further steps. An analysis rarely consists of data from single source, so these steps are executed for all data sets. Next, related data sets are combined so that they can be presented in a single view or fed into an algorithm. Because the data wrangling process can be very complex, a data expert or an automated data quality algorithm validates the data before it is published to the wider audience. Once the data receive approval from the data expert, it is made available to the users and notifications are sent, if necessary, to any downstream applications.

*The authors are indebted to Jacek Tokar, Michal Antkowiak, and David Dittmann of Procter & Gamble for providing this Analytics in Action.

It is good practice to always maintain a copy of the original raw data and data set produced at intermediate steps of the data wrangling process.

Raw data, which is sometimes called source data or primary data, is data that have not been processed or prepared for analysis. **Data wrangling** is the process of cleaning, transforming, and managing raw data so it is more reliable and can be more easily accessed and used for analysis. This process, which is also sometimes referred to as data munging or data remediation, involves transforming raw data that cannot easily be analyzed into a new data set that is designed specifically to support analysis.

Data wrangling activities include:

- Merging multiple data sources into a single data set
- Identifying missing values in data and either filling or deleting the record or observation
- Identifying erroneous values in data and replacing them with the correct values
- Identifying and deleting duplicate observations or records in data
- Identifying extreme values in data, determining whether they are legitimate, and taking appropriate action
- Subsetting data to facilitate analyses and comparisons of subgroups of the raw data

The data wrangling process can be manual or automated, or it can be a combination of manual and automated. The process can also be ad hoc or systematic. As the size, complexity, and importance of the raw data increases, an automated and systematic approach to data wrangling becomes more necessary. No matter how it is accomplished, the objective of data wrangling is to facilitate the efforts of data scientists and analytics professionals to focus on their analyses and the extraction of meaningful information from the data.

Although each data wrangling project is unique, the objective is to produce a final data set that is accurate, reliable, and accessible. This objective leads to a series of steps that are commonly taken in the data wrangling process:

- **Discovery:** The goal of the discovery phase is for the analyst to become familiar with the raw data in order to be able to conceptualize how it might be used. During this step, the analyst often discovers potential issues that will need to be addressed later in the data wrangling process. Just as data wrangling is a precursor to meaningful

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If new raw data are added to the existing raw data set in the enriching step, the new raw data should be subjected to the discovery, structuring, and cleaning steps. Also, it is good practice to always maintain a copy of the enriched raw data

analysis, discovery is a precursor to the data wrangling steps that follow, and this step should not be overlooked or neglected.

- **Structuring:** In the structuring phase, the analyst arranges the raw data so it can be more readily analyzed in the intended manner.
- **Cleaning:** The goal of the cleaning phase is to find and correct errors in the raw data that might distort the ensuing analyses. This can include addressing missing or incorrect values.
- **Enriching:** In this step, the analyst may augment the raw data by incorporating values from other data sets and/or applying transformations to portions of the existing data to ensure that all data required for the ensuing analyses will be included in the wrangled data set.
- **Validating:** In the validating phase, the analyst verifies that the wrangled data are accurate and reliable and that they are ready for the ensuing analyses.
- **Publishing:** In this step, the analyst creates a file containing the wrangled data and documentation of the file's contents, and makes these available to their intended users in a format they can use.

The quality of any analysis of data is limited by the quality of the data being analyzed. Data that are incomplete, inaccurate, unreliable, or inaccessible may limit the insights that can be gained from the analysis or even result in erroneous conclusions and poor decisions. By properly wrangling the raw data, the analyst reduces these risks by ensuring data are reliable and accessible before they are analyzed. Although data wrangling can be time-consuming and resource intensive, it is critical that organizations methodically and conscientiously follow the steps of the data wrangling process before moving on to the analysis of the data. Likewise, before data are used in more sophisticated models, accuracy of the data is of paramount importance. The old adage "garbage in = garbage out" rings true in analytics. In the ensuing sections of this chapter, we will discuss each of the steps of the data wrangling process in more detail.

4.1 Discovery

Data wrangling starts with discovery, as the analyst becomes familiar with, and understands, the raw data with an eye toward how the data should be organized to facilitate its use and analysis. The analyst should come to this step with a thorough understanding of how these data may be analyzed and consider this as the discovery process proceeds. This understanding includes how the raw data can be accessed, the structure of the raw data, whether the data are numerical or text, how they are arrayed or arranged, and the types of file(s) that contain the raw data.

In this section, we will consider issues that arise in the wrangling of raw data to ultimately be used in the creation and maintenance of a database for Stinson's MicroBrew Distributor, a licensed regional independent distributor of beer and a member of the National Beer Wholesalers Association. Stinson's provides refrigerated storage, transportation, and delivery of premium beers produced by several local microbreweries. The company's facilities include a state-of-the-art temperature-controlled warehouse and a fleet of temperature-controlled trucks. Stinson's employs sales, receiving, warehousing/inventory, and delivery personnel. When making a delivery, Stinson's monitors the retailer's shelves, taps, and keg lines to ensure the freshness and quality of the product. Because beer is perishable and because microbreweries often do not have the capacity to store, transport, and deliver large quantities of the products they produce, Stinson's holds a critical position in this supply chain.

Accessing Data

Accessing the data involves importing the raw data into whatever software is to be used to wrangle the raw data. Raw data can come in many formats, including:

- Delimited text (txt) files, comma separated value (csv) files, JavaScript Object Notation (JSON) files, XML files, Excel (xlsx) files, and HTML files
- Databases such as ACCESS, SQL, or ORACLE
- Web pages/HTML files

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Stinson's routinely receives a csv file that contains information on each delivery the company received from its supplier microbreweries during a given week. These data for one week, as stored in the csv file *sdeliveries.csv*, are shown in Figure 4.1. The names of the fields are shown in the first row of the data. In Figure 4.1, we see that the file *sdeliveries.csv* contains:

DATA *file*
sdeliveries.csv

SOrderNumber: Stinson's internal number assigned to each order placed with a microbrewery
BrewerID: Stinson's internal identification number assigned to the microbrewery with which the order has been placed
EmployeeID: The identification number of the Stinson's employee who received the delivery
Date of SDelivery: The date on which Stinson's received the delivery
SQuantity Delivered: The quantity (in units) received in the delivery

A csv file is a file that is in comma separated values format.

Each ensuing row in the file *sdeliveries.csv* represents one record or delivery, and tags or markers called **delimiters** indicate how to separate the raw data into various fields when they are read. Commas are used to delimit the fields in each record in the file *sdeliveries.csv*.

Many of the tools that can be used for data wrangling, such as Excel, R, Python, SQL, and Access, can import raw data from a csv file. Excel can automatically open and import a csv file, so if Stinson's wants to import the raw data into Excel, it can do so through the following steps once Excel is open:

Step 1. Click the **File** tab on the Ribbon and select **Open**

Figure 4.1 Display of the CSV File *sdeliveries* That Contains Information on Deliveries Received by Stinson's from Various Microbreweries During a Recent Week

```
SOrderNumber,BrewerID,EmployeeID,Date of SDelivery,SQuantity Delivered
17351,3,94,11/5/2023,3
17352,9,9,11/5/2023,6
17353,7,135,11/6/2023,2
17354,3,94,11/6/2023,3
17355,2,135,11/6/2023,2
17356,6,135,11/6/2023,5
17358,9,135,11/7/2023,3
17359,4,135,11/7/2023,2
17360,3,94,11/8/2023,8
17361,2,135,11/8/2023,1
17362,7,94,11/8/2023,2
17363,9,94,11/9/2023,4
17364,6,135,11/8/2023,2
17365,2,94,11/9/2023,5
17366,3,94,11/9/2023,4
17366,7,135,11/10/2023,4
17366,9,94,11/9/2023,4
17367,4,94,11/9/2023,3
```

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Step 2. Click on **Browse**, select **All Files (*)** in the **File Type** dropdown list in the **Open** dialog box, and navigate to the location of the file *tblpurchaseprices.txt*
Step 3. Select the file and click **Open**

Stinson's also maintains a txt file that contains the prices charged by each microbrewery that supplies Stinson's with beer. These data, as stored in the file *tblpurchaseprices.txt*, are shown in Figure 4.2. The names of the fields are shown in the first row of the data. In Figure 4.2, we see that the file *tblpurchaseprices.txt* contains:

BrewerID: Stinson's internal identification number assigned to the microbrewery with which the order has been placed
KegPurchasePrice: The price the microbrewery associated with the record charges for a keg of its beer
CasePurchasePrice: The price the microbrewery associated with the record charges for a case of its beer

DATA*file*
tblpurchaseprices.txt

Each ensuing row in the file *tblpurchaseprices.txt* represents one microbrewery, and the information in the records is tab delimited.

Many data wrangling software packages import delimited raw data that utilize tabs or other delimiters. If Stinson's wants to import the raw data in the file *tblpurchaseprices.txt* into Excel, it can do so through the following steps once Excel is open:

Step 1. Click the **File** tab on the Ribbon and select **Open**
Step 2. Click on **Browse**, select **All Files (*)** in the **File Type** dropdown list in the **Open** dialog box, and navigate to the location of the file *tblpurchaseprices.txt*
Step 3. Select the file and click **Open**
Step 4. In the **Text Import Wizard—Step 1 of 3** dialog box (shown in Figure 4.3)
Select **Delimited** in the **Original data type: Chose the file type that best describes your data** area
Indicate that the delimited raw data to be imported (*tblpurchases.txt*) has headers
Click **Next**

You can use the Text Import Wizard to import data from files that use other delimiters by indicating the appropriate delimiter in the **Text Import Wizard—Step 2 of 3** *dialog box.*

Step 5. In the **Text Import Wizard—Step 2 of 3** dialog box (shown in Figure 4.4)
Select **Tab** in the **Delimiters** area
Review the raw data in the **Data preview** area to ascertain that the data will import correctly
Click **Finish**

Figure 4.2 Display of File *tblpurchaseprices.txt* Containing Information About the Price Charged by Each Microbrewery That Supplied Stinson's with Beer During a Recent Week

BrewerID	KegPurchasePrice	CasePurchasePrice
1	$158.00	$43.00
2	$227.00	$62.00
3	$214.00	$59.00
4	$156.00	$43.00
5	$248.00	$68.00
6	$191.00	$52.00
7	$179.00	$49.00
8	$237.00	$65.00
9	$245.00	$67.00
10	$181.00	$50.00
11	$236.00	$65.00

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.3 Text Import Wizard—Step 1 of 3 Dialog Box

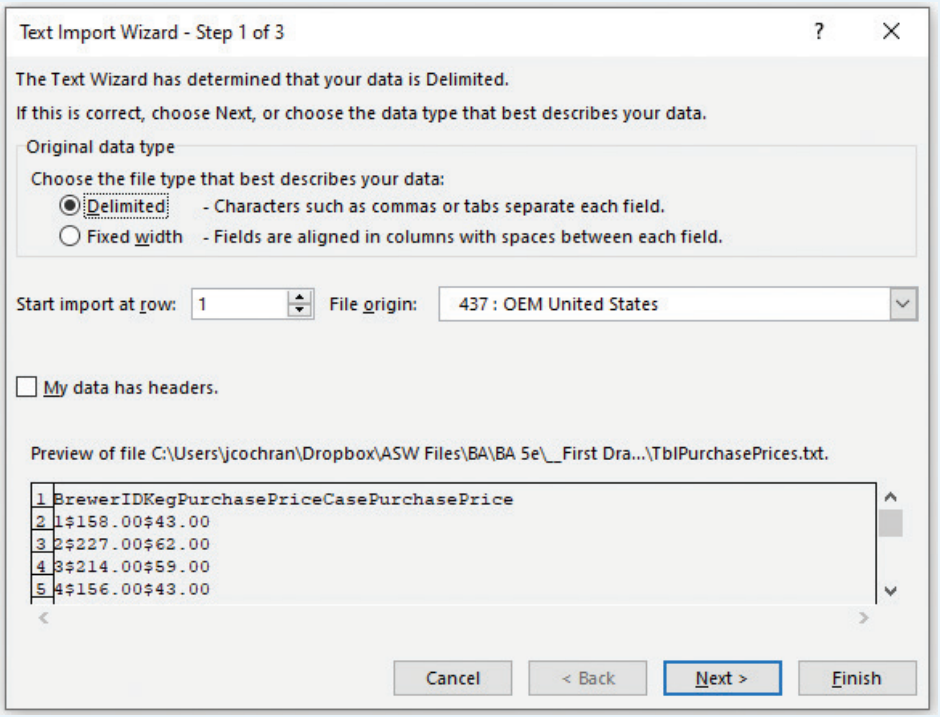
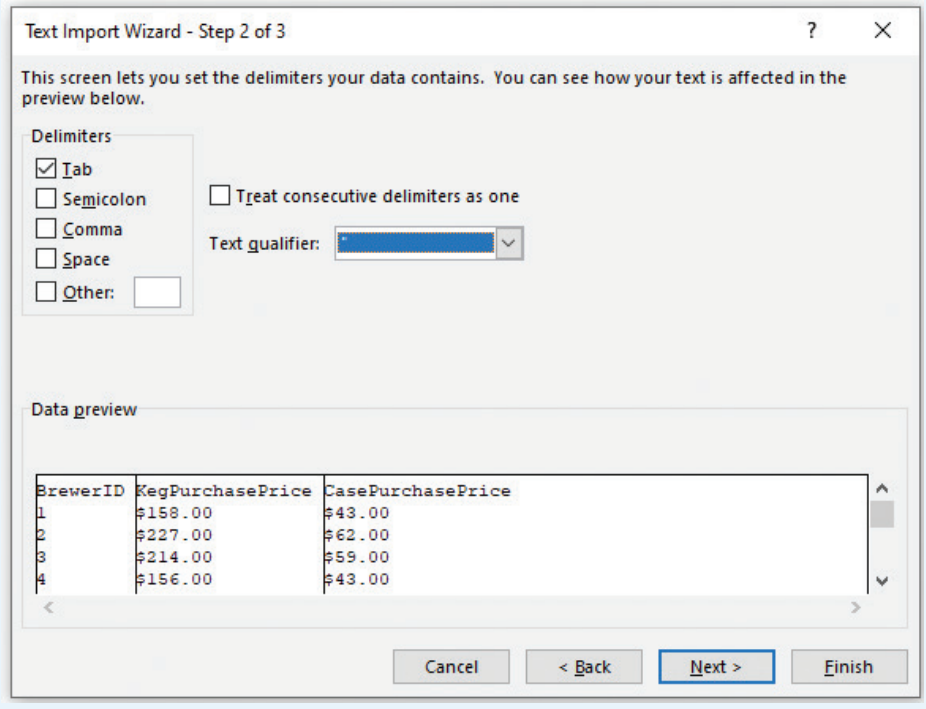


Figure 4.4 Text Import Wizard—Step 2 of 3 Dialog Box



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 4.1 Structured Data from the File tblbrewersaddresses

BrewerID	Brewery Name	Address	Phone Number
3	Oak Creek Brewery	12 Appleton St, Dayton, OH, 45455	937-449-1212
6	Gonzo Microbrew	1515 Marlboro Lane, Dayton, OH, 45429	937-288-2651
4	McBride's Pride	425 Dry River Rd, Miamisburg, OH, 45459	937-441-0123
9	Fine Pembrook Ale	141 Dusselberg Ave, Trotwood, OH, 45426	937-852-8752
7	Midwest Fiddler Crab	844 Small Hills Ave, Kettering, OH, 45453	937-633-7183
2	Herman's Killer Brew	912 Highlite Dr, Fairborn, OH, 45442	937-883-2651

The Format of the Raw Data

The format of the raw data determines how the data are wrangled. Each **record** in a data set is a grouping of characteristics for a particular observation, and each **field** is a characteristic of the observations in a data set. **Structured data** refers to data that are arrayed in a predetermined pattern that make them easy to manage and search. The most common pattern in which structured data are arrayed as a rectangle, with each row representing an observation or record and each column representing a variable or field. A file that is arrayed in this manner is often referred to as a **flat file**. Table 4.1 provides an example of structured data from the file *tblbrewersaddresses*, which contains contact information about each microbrewery that supplies Stinson's with beer.

The simple and orderly arrangement and formatting of structured data make them relatively easy to process. However, this also imposes constraints on what structured data can contain.

Unstructured data are databases that are not arranged in a predetermined pattern, and therefore in their raw form cannot be stored in a manner similar to that of a flat file. Text, multimedia, email messages, social media posts, chats, photographs, satellite images, medical scans, Internet of Things (IoT) sensor data, webpages, and video and audio files are common examples of files that contain unstructured data. Figure 4.5 provides an example of unstructured data from responses to an open-ended survey question dealing with why respondents do not purchase or use Ship-Shape Basin bathroom cleaner.

Because of its complex arrangement and formatting, raw unstructured data can be difficult to process and analyze Although unstructured data are more difficult to search and use than structured data, it is abundant. Experts estimate that over 80% of the world's data is unstructured. Unstructured data can also be useful, as it often contains unique data that can be harvested for deep and novel insights.

Figure 4.5 Unstructured Data from Responses to an Open-Ended Survey Question: "Why Respondents Do Not Purchase or Use Ship-Shape Basin Bathroom Cleaner?"

Ship-Shape smells terrible—like a doctor's office—and the smell lingers for days after I have used it.

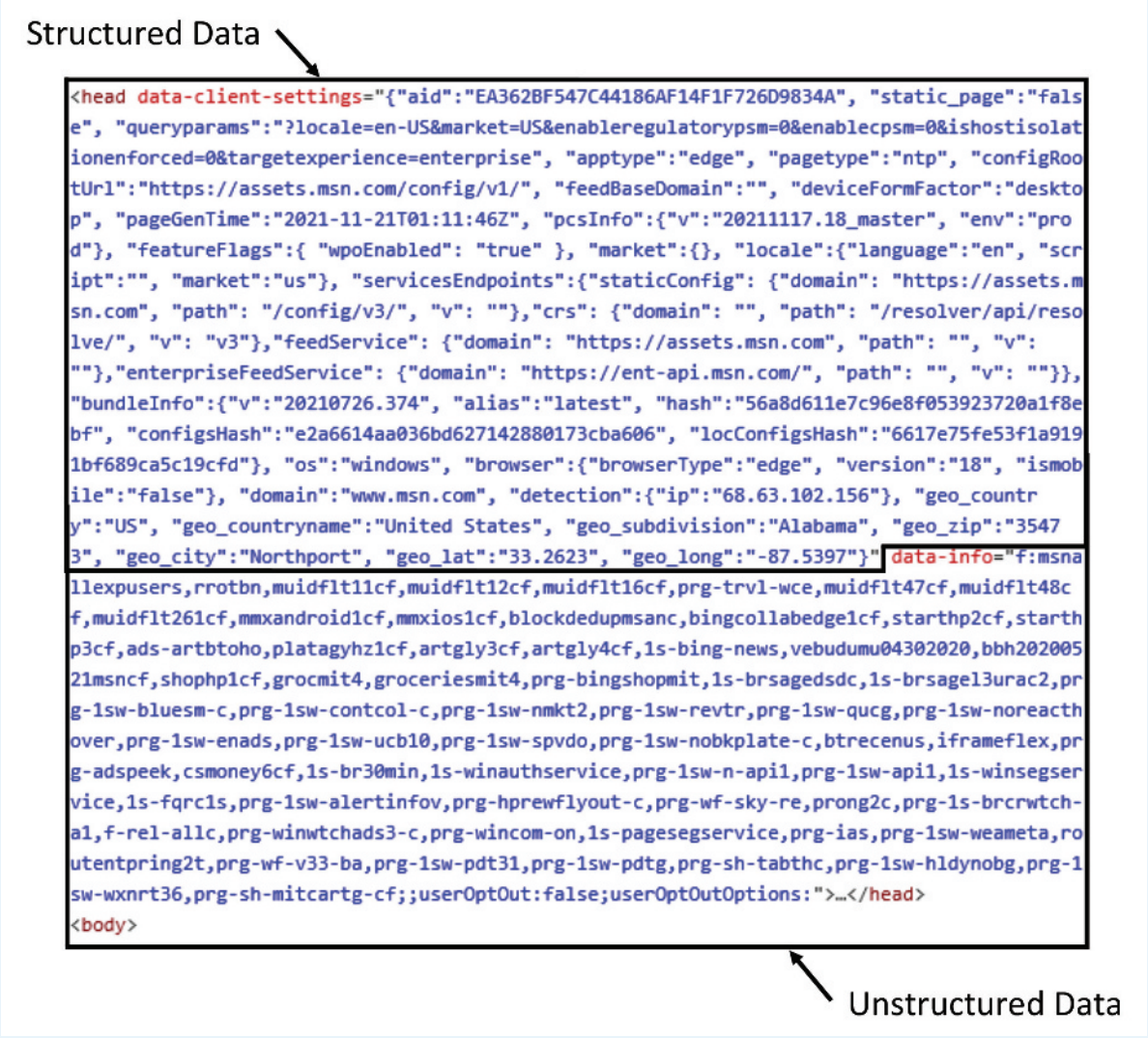
The bright neon lime green color of the foam is nauseating.

The smell is extremely mediciney—it gives me a headache. And the product doesn't get my sinks and toilets very clean—it is too much work to clean with Ship-Shape.

Ship-Shape leaves big streaks on whatever I use it to clean—then I have to use something else to clean the streaks!!!

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.6 Semi-Structured Data from an Internet Browsing Session in Microsoft Edge



Semi-structured data does not have the same level of organization as structured data, but it does contain elements that allow for isolation of some elements of the raw data when they are imported. An example of a semi-structured file is the record of an internet browsing session shown in Figure 4.6. Some elements of the file do not have a pre-defined structure, but other elements (such as "locale," "apptype," "os," and "pageGenTime") have structure that allows them to be easily identified and isolated.

Raw data in each of these formats can be valuable, and businesses commonly collect and need to analyze structured, unstructured, and semi-structured data.

4.2 Structuring

Once the analyst understands the raw data and how it should be organized to facilitate its analysis, the focus turns to structuring the raw data file so it can be more readily analyzed in the intended manner. This may include the formatting of the data fields, how the data are to be arranged, splitting one field with several important pieces of information into several fields, and merging several fields into a single more meaningful field.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Data Formatting

The two primary forms of data are numeric and text. **Numeric data** are data that are stored in a manner that allows mathematical operations to be performed on them. Data of this type generally represent a count or measurement.

Software can often be used to apply specific formats to fields in raw data. For example, Excel offers several formats that can be applied to numerical data, including Currency, Accounting, Long Date, Short Date, Time, Percentage, and Scientific Notation. For example, we may want to use a different format for the date field in the csv file *sdeliveries.csv* (shown in Figure 4.1). To change the format of this field in Excel, we apply the following steps after importing the file *sdeliveries.csv* into Excel:

Step 1. Select cells D2:D19 (the range of cells to format)
Step 2. Right-click one of the highlighted cells and select **Format Cells . . .**
Step 3. When the **Format Cells** dialog box opens:
Select **Date** in the **Category** pane and click **March 14, 2012** (the desired format for this field) in the **Type** pane as shown in Figure 4.7
Click **OK**

The data in the Date field are now shown in the new format.

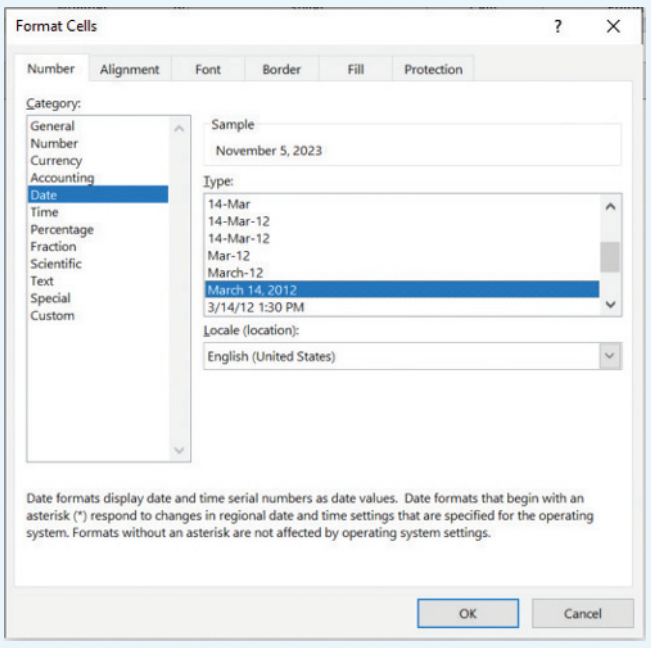
DATA*file*
sdeliveries.csv

Text data are data that are words, phrases, sentences, and paragraphs. This kind of data includes characters such as alphabetical, numerical (when the numbers are not to be used in mathematical calculations, such as with a field for which 1 corresponds to YES and 2 corresponds to NO), and special symbols. Brewery Name and Address in the file *tblbrewersaddresses.csv* (shown in Figure 4.2) are examples of text data.

Arrangement of Data

We have considered data sets that are arrayed in a manner such that every row (record) represents an observation on a set of variables (fields), and the values of those variables

Figure 4.7 The Excel Format Cells Dialog Box



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

are listed in the columns. However, there are multiple ways to organize the data within this structure, particularly if the data contain categorical variables.

Consider data on three-point shot attempts by each team in a high school basketball conference for the past six years. Figure 4.8a and Figure 4.8b show the same portion of these data arranged in two different ways. In Figure 4.8a, each record (row) consists of values in three fields (columns): a team name, a year, and the number of three-point shot attempts by that team during that year. In Figure 4.8b, each record (row) consists of values in seven fields (columns): a team name, and then in the next six columns, the number of three-point shot attempts by that team in each of six years.

Figure 4.8 A Portion of Data from the File hs3pa in Stacked (Panel a) and Unstacked (Panel b) Form

(a) Stacked Data

	A	B	C
1	**Team**	**Year**	**Three Point Attempts**
2	Gamaliel Gremlins	1	602
3	Fitzgerald Hawks	1	713
4	Delano Demons	1	647
5	Ulysses Gazelles	1	711
6	Milhous Rams	1	800
7	Birchard Spiders	1	682
8	Quincy Bobcats	1	800
9	Knox Golden Bears	1	650
10	Gamaliel Gremlins	2	918
11	Fitzgerald Hawks	2	755
12	Delano Demons	2	759
13	Ulysses Gazelles	2	873
14	Milhous Rams	2	890
15	Birchard Spiders	2	762
16	Quincy Bobcats	2	631
17	Knox Golden Bears	2	794
18	Gamaliel Gremlins	3	776
19	Fitzgerald Hawks	3	808
20	Delano Demons	3	653
21	Ulysses Gazelles	3	919

(b) Unstacked Data

	A	B	C	D	E	F	G
1		**Three Point Attempts**					
2	**Team**	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Year 6**
3	Gamaliel Gremlins	602	918	776	590	600	559
4	Fitzgerald Hawks	713	755	808	841	831	628
5	Delano Demons	647	759	653	779	626	826
6	Ulysses Gazelles	711	873	919	676	775	645
7	Milhous Rams	800	890	506	695	775	764
8	Birchard Spiders	682	762	587	800	903	773
9	Quincy Bobcats	800	631	626	910	894	670
10	Knox Golden Bears	650	794	737	772	642	717

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The arrangement of data in Figure 4.8a is referred to as stacked data. **Stacked data** are data organized so that the values for each variable are stored in a single field. The arrangement of data in Figure 4.8b is referred to as unstacked data. **Unstacked data** are data organized so that the values for one variable correspond to separate fields. The data in Figure 4.8a are stacked with respect to Year because all values of Year are listed in a single field; each record corresponds to a unique combination of team and year. The data in Figure 4.8b are unstacked with respect to Year because there is a field corresponding to each value of Year; each record corresponds to a single team for each of several years. The same data are contained in both versions, but the data are arranged differently.

Depending on the source of the data and how it was collected, data may be provided in stacked or unstacked form. Both arrangements of data can be useful as each arrangement can facilitate different analyses. For example, the unstacked version of the raw data facilitates the analysis of individual team performances across multiple years, and the stacked version of the raw data facilitates the analysis of the performances in individual years across multiple teams. Some software packages also require the data be in stacked or unstacked form, so it is useful to be able to transform stacked data to unstacked data and transform unstacked data to stacked data. In the following steps, we use Excel to convert the three-point shot attempt data from stacked form in the file *stackedtablehs3pa* (shown in Figure 4.8a) to unstacked form with respect to the Year variable.

DATA *file*
stackedtablehs3pa

Step 1. Select any cell in the range of the Table, for example, cell A1
Step 2. Click the **Data** tab in the Ribbon
Step 3. In the **Get & Transform** group, click **From Table/Range**
Step 4. When the **Power Query Editor** window appears:
Select the **Year** column
Click the **Transform** tab
In the **Any Column** group, click **Pivot Column**
When the **Pivot Column** dialog box appears, select **Three Point Attempts** in the drop-down menu for **Values Column**
Click **OK**
Click the **Home** tab and in the **Close** group, click **Close & Load**

The preceding steps create an Excel Table that contains the same fields and records in a new worksheet that are arranged in the same manner as the unstacked data form in Figure 4.8b.

Conversely, in the following steps we convert the three-point shot attempt data from unstacked form in the file *unstackedtablehs3pa* (shown in Figure 4.8b) to stacked form with respect to the Year variable:

DATA *file*
unstackedtablehs3pa

Step 1. Select any cell in the range of the Table, for example, cell A3
Step 2. Click the **Data** tab in the Ribbon
Step 3. In the **Get & Transform** group, click **From Table/Range**
Step 4. When the **Power Query Editor** window appears
Select the columns corresponding to **Year1**, **Year 2**, **Year 3**, **Year 4**, **Year 5**, and **Year 6** (press the **Ctrl** key while selecting each column)
Click the **Transform** tab
In the **Any Column** group, click **Unpivot Columns**
Click the **Home** tab and in the **Close** group, click **Close & Load**

The preceding steps create an Excel Table in a new worksheet that after sorting on years, matches the stacked data form in Figure 4.8a.

Splitting a Single Field into Multiple Fields

Sometimes one field in the raw data contains several pieces of information that would be more useful if divided into separate fields. For example, dates, addresses, and times often contain multiple pieces of information that may be more useful if divided into separate

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

fields. Consider the Address field in the file *tblbrewersaddresses* shown in Figure 4.2. The information contained in this field might be easier to use if the street address, city, state, and zip code were each in its own field. Such an arrangement would allow us to sort on City or Zip Code and analyze the data in that manner.

Many data wrangling software packages provide tools for splitting a single field into multiple fields. In Excel, on the **Data** tab in the **Data Tools** group, the **Text to Columns** button invokes the **Convert Text to Columns Wizard** that can be used for this purpose. Useful operations that can be performed using the **Convert Text to Columns Wizard** include:

- Splitting names into the first name and last name
- Splitting email addresses into username and domain name
- Extracting the first *n* characters of a string

We will illustrate splitting one field into multiple fields by applying the **Convert Text to Columns Wizard** tool to raw data in the file *tblbrewersaddresses*. We will separate the field Address as shown in Figure 4.9 into the individual fields Street Address, City, State, and Zip Code.

DATA *file*
tblbrewersaddresses

Once the file *tblbrewersaddresses* is open:

Step 1. Select columns D, E, and F by clicking on the heading of column D and dragging the cursor across the headings of columns E and F

Step 2. Right click anywhere in the selected columns and select **Insert** to insert three columns immediately to the right of the Address column to provide space for the new columns to be generated (the four columns generated for the new Address, City, State, and Zip Code fields will fill the column now occupied by the Address field and the three new columns)

Step 3. Select cells C2:C7 (the range of the field to be separated)

Step 4. Click the **Data** tab in the Ribbon

Step 5. In the **Data Tools** group, click **Text to Columns**

Step 6. When the **Convert Text to Columns Wizard—Step 1 of 3** dialog box appears (as shown in Figure 4.10)
Select **Delimited** (because the data in the Address field that is to be put into individual columns is separated by commas)
Click **Next**

Step 7. When the **Convert Text to Columns Wizard—Step 2 of 3** dialog box appears (as shown in Figure 4.11)
In the **Delimiters** pane, select **Comma** and deselect all other delimiter options
Click **Next**

Figure 4.9 Data From the File *tblbrewersaddresses*

	A	B	C	D
1	BrewerID	Brewery Name	Address	Phone Number
2	3	Oak Creek Brewery	12 Appleton St, Dayton, OH, 45455	937-449-1212
3	6	Gonzo Microbrew	1515 Marlboro Lane, Dayton, OH, 45429	937-288-2651
4	4	McBride's Pride	425 Dry River Rd, Miamisburg, OH, 45459	937-441-0123
5	9	Fine Pembrook Ale	141 Dusselberg Ave, Trotwood, OH, 4542	937-852-8752
6	7	Midwest Fiddler Crab	844 Small Hills Ave, Kettering, OH, 45453	937-633-7183
7	2	Herman's Killer Brew	912 Highlite Dr, Fairborn, OH, 45442	937-883-2651

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.10 Convert Text to Columns Wizard—Step 1 of 3 Dialog Box

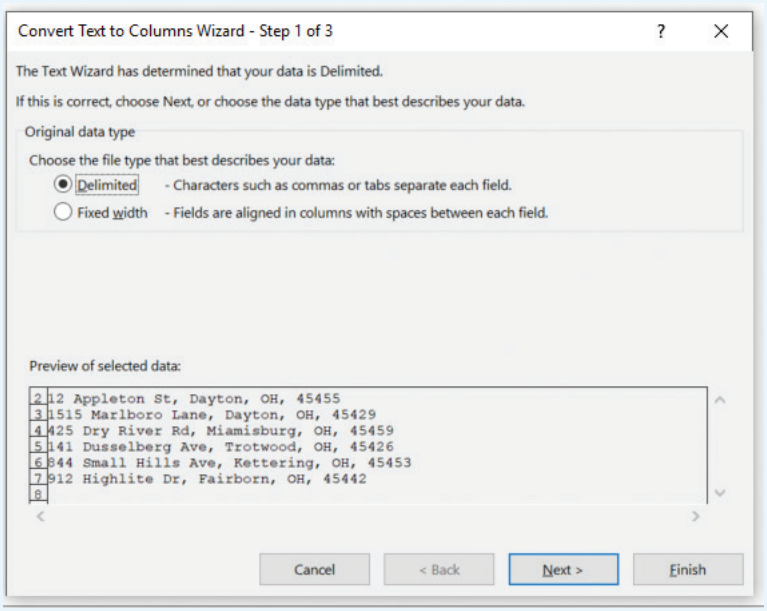
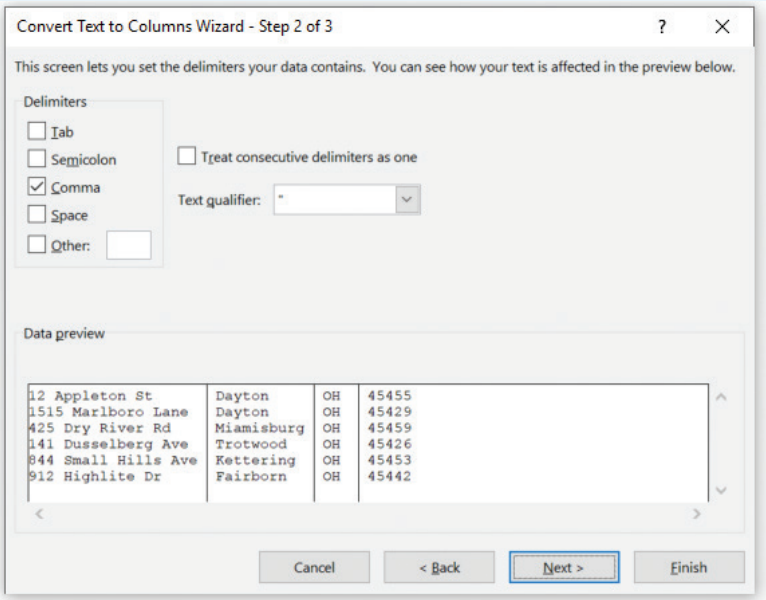
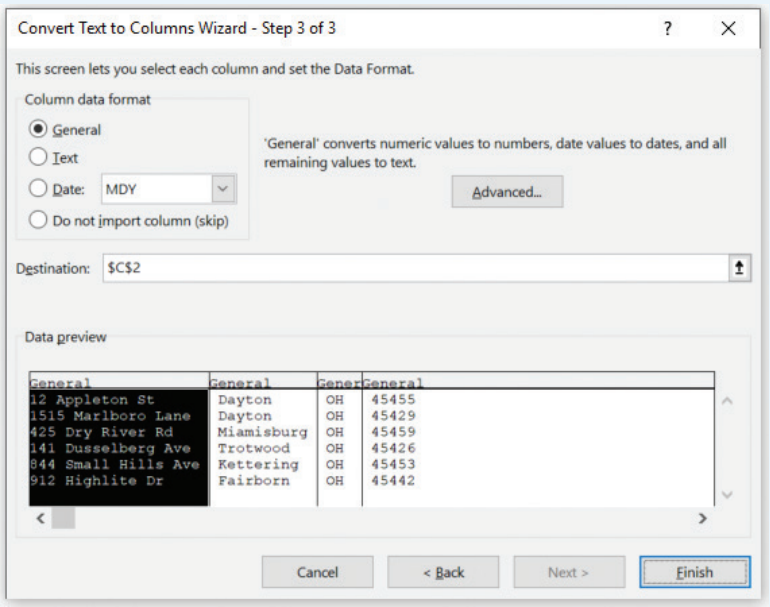


Figure 4.11 Convert Text to Columns Wizard—Step 2 of 3 Dialog Box



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.12 Convert Text to Columns Wizard – Step 3 of 3 Dialog Box



Step 8. When the **Convert Text to Columns Wizard—Step 3 of 3** dialog box appears (as shown in Figure 4.12)

In the **Column data format** pane, click the first column in the **Data preview** area
Select **General** in the **Column data format** area
Repeat this process for each column
Click **Finish**

We have created the worksheet in Figure 4.13.

We can now add appropriate headers to columns C, D, E, and F and save the file under the name *tblbrewers*.

Figure 4.13 Contents of the File *tblbrewersaddresses* with the Address Field Split into Four Fields

	A	B	C	D	E	F	G
1	**BrewerID**	**Brewery Name**	**Address**				**Phone Number**
2	3	Oak Creek Brewery	12 Appleton St	Dayton	OH	45455	937-445-1212
3	6	Gonzo Microbrew	1515 Main St	Dayton	OH	45429	937-278-2651
4	4	McBride's Pride	425 Mad River Rd	Miamisburg	OH	45459	937-439-0123
5	9	Fine Pembrook Ale	141 Dusselberg Ave	Trotwood	OH	45426	937-837-8752
6	7	Midwest Fiddler Crab	844 Far Hills Ave	Kettering	OH	45453	937-633-7183
7	2	Herman's Killer Brew	912 Airline Dr	Fairborn	OH	45442	937-878-2651

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.14 Excel Spreadsheet Showing the Names Data

	A	B	C	D	E
1	**Prefix**	**First Name**	**Middle Name**	**Last Name**	**Suffix**
2	Dr.	Zuri	G	Leseoana	
3	Ms.	Donna		Spicoli	
4	Mr.	Juan	Mateo	Rivera	Jr.
5		Donny	John	Moroni	
6	Mr.	Albert		Levy	III
7	Mr.	Haneul		Seong	
8		Meera		Bhatt	
9	Ms.	Jasmine	D	Belcher	

Combining Multiple Fields into a Single Field

Sometimes several fields in the raw data contain information that would be more useful if combined into a single field. For example, if we received the data shown in Figure 4.14 but needed the prefix, first name, middle name, and last name all in a single field (perhaps to facilitate the creation of shipping labels), we would want to combine these four fields into a single field.

After entering Dr. Zuri G Leseoana *in cell F2, you can also use the Flash Fill functionality to automatically fill the remaining records in the column F by* clicking the **Data** tab in the Ribbon, then *clicking* **Flash Fill** *In the* **Data Tools** *group.*

Many data wrangling software packages provide tools for combining multiple fields into a single field. In Excel, the Flash Fill functionality will quickly combine multiple fields into a single field. If we enter *Dr. Zuri G Leseoana* in cell F2, and then start typing *Ms.* in cell F3, Excel will fill in column F for each ensuing record. Figure 4.15a shows the Flash Fill function filling the remainder of cell F3 as well as all of cells C4:C7.

Note that Flash Fill replicates the pattern it observes in the first entry, so for each record it will add a blank space for each missing field in ensuing records. Thus, for the Names data Flash Fill adds a blank space in front of the first name for the entries it creates for Donny John Moroni and Meera Bhatt (because these records do not include a prefix). Flash Fill also adds a blank space between the first name and last name in the entries it creates for Ms. Donna Spicoli, Mr. Albert Levy, and Mr. Haneul Seong (because these records do not include a middle name or initial). Also note that Flash Fill will not pick up the suffixes in the Names data because there is no suffix in the entry in cell F2 (the cell that was used to establish the pattern to be replicated).

Figure 4.15a Excel Flash Fill Function Filling the Remainder of Cell F3 and All of Cells C4:C7

	A	B	C	D	E	F	G
1	**Prefix**	**First Name**	**Middle Name**	**Last Name**	**Suffix**		
2	Dr.	Zuri	G	Leseoana		Dr. Zuri G Leseoana	
3	Ms.	Donna		Spicoli		Ms. Donna Spicoli	
4	Mr.	Juan	Mateo	Rivera	Jr.	Mr. Juan M	
5		Donny	John	Moroni		Donny Joh	
6	Mr.	Albert		Levy	III	Mr. Albert	
7	Mr.	Haneul		Seong		Mr. Haneu	
8		Meera		Bhatt		Meera Bh	
9	Ms.	Jasmine	D	Belcher		Ms. Jasmi	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.15b Excel Spreadsheet Showing the Names Data with the Fields Prefix, First Name, Middle Name, Last Name, and Suffix Combined in a Single Field

	A	B	C	D	E	F	G
1	**Prefix**	**First Name**	**Middle Name**	**Last Name**	**Suffix**		
2	Dr.	Zuri	G	Leseoana		Dr. Zuri G Leseoana	Dr. Zuri G Leseoana
3	Ms.	Donna		Spicoli		Ms. Donna Spicoli	Ms. Donna Spicoli
4	Mr.	Juan	Mateo	Rivera	Jr.	Mr. Juan Mateo Rivera	Mr. Juan Mateo Rivera, Jr.
5		Donny	John	Moroni		Donny John Moroni	Donny John Moroni
6	Mr.	Albert		Levy	III	Mr. Albert Levy	Mr. Albert Levy, III
7	Mr.	Haneul		Seong		Mr. Haneul Seong	Mr. Haneul Seong
8		Meera		Bhatt		Meera Bhatt	Meera Bhatt
9	Ms.	Jasmine	D	Belcher		Ms. Jasmine D Belcher	Ms. Jasmine D Belcher

We can resolve these problems and combine the fields Prefix, First Name, Middle Name, Last Name, and Suffix into a single field for each record with Excel's TEXTJOIN function. We illustrate combining multiple fields into a single field by applying the TEXTJOIN function to the raw data in the file *names* (shown in Figure 4.14) to create a single field containing the contents of the Prefix, First Name, Middle Name, Last Name, and Suffix fields.

Because we want to delimit the fields Prefix, First Name, Middle Name, and Last Name with spaces and delimit the combination of these fields and the field Suffix with a comma in the final result, we will work in two steps:

DATA*file*
names

Step 1. In cell F2, enter *=TEXTJOIN(" ",TRUE,A2:D2)* where the first argument " " of this function defines the delimiter (a space in this example), the second argument of "TRUE" indicates that empty cells should be ignored, and the third argument identifies the cells to be joined

Step 2. Copy cell F2 into cells F3:F9

Step 3. In cell G2, enter *=TEXTJOIN(", ",TRUE,F2,E2)* where the first argument "," of this function defines the delimiter (a comma in this example), the second argument "TRUE" indicates that empty cells should be ignored, and the third argument identifies the cells to be joined

Step 4. Copy cell G2 into cells G3:G9

The results of these steps appear in Figure 4.15b. Using Excel's. TEXTJOIN function, we have created a new field in column G that contains the contents of the fields Prefix, First Name, Middle Name, Last Name, and Suffix for each record with no extraneous blank spaces.

Notes + Comments

1. In addition to numeric and text data, Excel also supports logical data, which returns the value TRUE or FALSE as the product of a test or comparison, and error data, which occurs when Excel recognizes a mistake or missing information while processing an entry.
2. Excel's CONCAT function can also be used to combine the contents of multiple fields into a single field. As with Flash Fill, the CONCAT function does not ignore empty fields. However, the CONCAT function does allow the user to control the fields to be combined.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4.3 Cleaning

In Chapters 2 and 3, we discuss many descriptive statistics and data-visualization methods that are useful for gaining an understanding of data.

Once the discovery and structuring phases of data wrangling are complete, the analyst moves on to cleaning the raw data. This step includes identifying missing data, erroneous data, duplicate records, and outliers and determining the best actions for addressing these issues. Data cleaning makes heavy use of descriptive statistics and data-visualization methods to identify missing data, erroneous data, and outliers.

Missing Data

Data sets commonly include observations with missing values for one or more variables. In some cases, there is an appropriate reason for a value of a field to be missing; these are called **legitimately missing data**. For example, respondents to a survey may be asked if they belong to a fraternity or a sorority, and then in the next question are asked how long they have belonged to a fraternity or a sorority. If a respondent does not belong to a fraternity or a sorority, they should skip the ensuing question about how long. Generally, no remedial action is taken for legitimately missing data.

In other cases, there is not an appropriate reason for the value of a field to be missing; these are called **illegitimately missing data**. These cases can result for a variety of reasons, such as a respondent electing not to answer a question that they are expected to answer, a respondent dropping out of a study before its completion, or sensors or other electronic data collection equipment failing during a study. Remedial action is considered for illegitimately missing data. If possible, the original source of the data should be checked for all instances of illegitimately missing data to establish whether the values that are missing can be determined. If the missing value can be found in the source of the data, the verified value from the source of the data should be used to replace the missing value. The primary options for addressing missing data if the missing value cannot be found at its source are: (1) to discard records (rows) with any missing values, (2) to discard any field with missing values, (3) to fill in missing entries with estimated values, or (4) to apply a data-mining algorithm (such as classification and regression trees) that can handle missing values.

Regression trees are discussed in Chapter 10. Classification trees are discussed in Chapter 11.

Deciding on a strategy for dealing with missing data requires some understanding of why the data are missing and the potential impact these missing values might have on an analysis. If the tendency for an observation to be missing the value for some variable is entirely random, then whether data are missing does not depend on either the value of the missing data or the value of any other variable in the data. In such cases the missing value is called **missing completely at random** (MCAR). For example, if a missing value for a question on a survey is completely unrelated to the value that is missing and is also completely unrelated to the value of any other question on the survey, the missing value is MCAR.

However, the occurrence of some missing values may not be completely at random. If the tendency for an observation to be missing a value for some variable is related to the value of some other variable(s) in the data, the missing value is called **missing at random** (MAR). For data that are MAR, the reason for the missing values may determine its importance. For example, if the responses to one survey question collected by a specific employee were lost due to a data entry error, then the treatment of the missing data may be less critical. However, in a health care study, suppose observations corresponding to patient visits are missing the results of diagnostic tests whenever the doctor deems the patient too sick to undergo the procedure. In this case, the absence of a variable measurement provides additional information about the patient's condition, which may aid in understanding other relationships in the data.

A third category of missing data is **missing not at random** (MNAR). Data is MNAR if the tendency for the value of a variable to be missing is related to the value that is missing. For example, survey respondents with high incomes may be less inclined than respondents with lower incomes to respond to the question on annual income, and so these missing data for annual income are MNAR.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Understanding which of these three categories—MCAR, MAR, and MNAR—missing values fall into is critical in determining how to handle missing data. If a variable has observations for which the missing values are MCAR or MAR and only a relatively small number of observations are missing values, the observations that are missing values can be ignored. We will certainly lose information if the observations that are missing values for the variable are ignored, but the results of an analysis of the data will not be biased by the missing values.

If a variable has observations for which the missing values are MNAR, the observation with missing values cannot be ignored because any analysis that includes the variable with MNAR values will be biased. If the variable with MNAR values is thought to be redundant with another variable in the data for which there are few or no missing values, removing the MNAR variable from consideration may be an option. For example, if the MNAR variable is highly correlated with another variable that is known for a majority of observations, the loss of information may be minimal.

Whether the missing values are MCAR, MAR, or MNAR, the first course of action when faced with missing values is to try to determine the actual value that is missing by examining the source of the data or logically determining the likely value that is missing. If the missing values cannot be determined and ignoring missing values or removing a variable with missing values from consideration is not an option, **imputation** (systematic replacement of missing values with values that seem reasonable) may be useful. Options for replacing the missing entries for a variable include replacing the missing value with the variable's mode, mean, or median. Imputing values in this manner is truly valid only if variable values are MCAR; otherwise, we may be introducing misleading information into the data. If missing values are particularly troublesome and MAR, it may be possible to build a model to predict a variable with missing values and then to use these predictions in place of the missing entries. How to deal with missing values is subjective, and caution must be used to not induce bias by replacing missing values.

As an example of how to deal with missing data, let us consider Blakely Tires, a U.S. producer of automobile tires. To learn about the conditions of its tires on automobiles in Texas, the company has obtained information for each of the four tires from 116 automobiles with Blakely brand tires that have been collected through recent state automobile inspection facilities in Texas. The data obtained by Blakely includes the position of the tire on the automobile (left front, left rear, right front, right rear), age of the tire, mileage on the tire, and depth of the remaining tread on the tire. Before Blakely management attempts to learn more about its tires on automobiles in Texas, it wants to assess the quality of these data.

The tread depth of a tire is a vertical measurement between the top of the tread rubber to the bottom of the tire's deepest grooves and is measured in 32nds of an inch in the United States. New Blakely brand tires have a tread depth of 10/32nds of an inch, and a tire's tread depth is considered insufficient if it is 2/32nds of an inch or less. Shallow tread depth is dangerous as it results in poor traction that makes steering the automobile more difficult. Blakely's tires generally last for four to five years or 40,000 to 60,000 miles.

We begin assessing the quality of these data by determining which (if any) observations have missing values for any of the variables in the *treadwear* data. We can do so using Excel's COUNTBLANK function. After opening the file *treadwear,* we follow the steps below.

DATA*file*
treadwear

Step 1. Enter the heading *# of Missing Values* in cell G2
Step 2. Enter the heading *Life of Tire (Months)* in cell H1
Step 3. Enter *=COUNTBLANK(C2:C457)* in cell H2

The result in cell H2 shows that none of the observations in these data is missing its value for Life of Tire.

By repeating this process for the remaining quantitative variables in the data (Tread Depth and Miles) in columns I and J, we determine that there are no missing values for Tread Depth and one missing value for Miles. The first few rows of the resulting Excel spreadsheet are provided in Figure 4.16.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.16 Portion of Excel Spreadsheet Showing Number of Missing Values for Variables in the Blakely Tires Data

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)		
2	13391487	LR	58.4	2.2	2805		# of Missing Values	0		
3	21678308	LR	17.3	8.3	39371					
4	18414311	RR	16.5	8.6	13367					
5	19778103	RR	8.2	9.8	1931					
6	16355454	RR	13.7	8.9	23992					
7	8952817	LR	52.8	3.0	48961					
8	6559652	RR	14.7	8.8	4585					
9	16289814	LR	6.2	10.1	5221					

Sorting data in Excel is discussed in Chapter 2.

Next we sort all of Blakely's data (cells A2:E457) on Miles from smallest to largest value to determine which observation is missing its value of this variable. When sorting data, Excel's sort procedure will list all observations with missing values for the sort variable in the last rows of the sorted data. Thus, the observations that are missing values for Miles will be the last observations in the *treadwear* data sorted on Miles as shown in Figure 4.17.

We can see in Figure 4.17 that the value of Miles is missing from the left front tire of the automobile with ID Number 3354942. Because only one of the 456 observations is missing its value for Miles, this is likely MCAR and so ignoring the observation would not likely bias any analysis we wish to undertake with these data. However, we may be able to salvage this observation by logically determining a reasonable value to substitute for this missing value. It is sensible to suspect that the value of Miles for the left front tire of the automobile with the ID Number 3354942 would be identical to the value of miles for the other three tires on this automobile, so we sort all the data on ID number and scroll through the data to find the four tires that belong to the automobile with the ID Number 3354942.

Occasionally missing values in a data set are indicated with a unique value, such as 9,999,999. Be sure to check to see if a unique value is being used to indicate a missing value in the data.

Figure 4.18 shows that the value of Miles for the other three tires on the automobile with the ID Number 3354942 is 33,254, so this may be a reasonable value for the Miles of the left front tire of the automobile with the ID Number 3354942. However, before substituting this value for the missing value of the left front tire of the automobile with ID Number 3354942, we should attempt to ascertain (if possible) that this value is valid—there are legitimate reasons why a driver might replace a single tire. In this instance we will assume that the correct value of Miles for the left front tire on the automobile with the ID Number 3354942 is 33,254 and substitute that number in the appropriate cell of the spreadsheet.

Figure 4.17 Portion of Excel Spreadsheet Showing the Blakely Tires Data Sorted on Miles in Ascending Order (with rows 5–454 hidden)

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)		
2	15890813	LF	16.1	8.6	206		# of Missing Values	0		
3	15890813	LR	16.1	8.6	206					
4	15890813	RF	16.1	8.6	206					
455	9306585	RF	45.4	4.2	107237					
456	9306585	RR	45.4	4.1	107237					
457	3354942	LF	21.4	8.5						

Note that rows 5-454 are hidden in this figure.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.18 Portion of Excel Spreadsheet Showing Blakely Tires Data Sorted in Ascending Order by ID Number

54	3121851	LR	17.1	8.4	21378
55	3121851	RR	17.1	8.4	21378
56	3121851	RF	17.1	8.4	21378
57	3121851	LF	17.1	8.5	21378
58	3354942	LF	21.4	7.7	
59	3354942	RF	21.4	7.7	33254
60	3354942	RR	21.4	7.8	33254
61	3354942	LR	21.4	7.7	33254
62	3574739	RR	73.3	0.2	57313
63	3574739	RF	73.3	0.2	57313
64	3574739	LF	73.3	0.2	57313
65	3574739	LR	73.3	0.2	57313

Identification of Erroneous Outliers, Other Erroneous Values, and Duplicate Records

Summary statistics, frequency distributions, bar charts and histograms, z-scores, scatter charts, correlation coefficients, and other tools that can be used to uncover data-quality issues and outliers are discussed in Chapters 2 and 3.

Examining the variables in the data set by use of summary statistics, frequency distributions, bar charts and histograms, z-scores, scatter charts, correlation coefficients, and other tools can uncover data-quality issues and outliers. For example, finding the minimum or maximum value for Tread Depth in the Blakely Tires data may reveal unrealistic values—perhaps even negative values—for Tread Depth, which would indicate a problem for the value of Tread Depth for any such record.

It is important to note here that many software packages, including Excel, ignore missing values when calculating various summary statistics such as the mean, standard deviation, minimum, and maximum. However, if missing values in a data set are indicated with a unique value (such as 9999999), these values may be included by software packages in calculations of various summary statistics such as the mean, standard deviation, minimum, and maximum. Both cases can result in misleading values for summary statistics, which is why many analysts prefer to deal with missing data issues prior to using summary statistics to attempt to identify erroneous outliers and other erroneous values in the data.

We again consider the Blakely tire data. We calculate the mean and standard deviation of each variable (age of the tire, mileage on the tire, and depth of the remaining tread on the tire) to assess whether values of these variable are reasonable in general. To see how to do this in Excel, return to the file *treadwear* and complete the following steps:

Step 1. Enter the heading *Mean* in cell G2
Step 2. Enter the heading *Standard Deviation* in cell G3
Step 3. Enter *Life of Tire (Months)* in cell H1
Enter *Tread Depth* in cell I1
Enter *Miles* in cell J1

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Step 4. Enter =*AVERAGE(C2:C457)* in cell H2
Enter =*AVERAGE(D2:E457)* in cell I2
Enter =*AVERAGE(E2:E457)* in cell J2
Step 5. Enter =*STDEV.S(C2:C457)* in cell H3
Enter =*STDEV.S(D2:D457)* in cell I3
Enter =*STDEV.S(E2:E457)* in cell J3

If you do not have good information on what are reasonable values for a variable, you can use z-scores to identify outliers to be investigated. z-scores are discussed in Chapter 2.

Figure 4.19 shows the results of these steps. The results in cells H2 and H3 show that the mean and standard deviation for life of tires are 23.8 months and 31.83 months, respectively. These values appear to be reasonable for the life of tires in months. The results for Tread Depth and Miles in columns I and J show that the mean and standard deviation for tread depth are 7.62/32nds of an inch and 2.47/32nds of an inch, respectively, and the mean and standard deviation for miles are 25,440.22 and 23,600.21, respectively. These values also appear to be reasonable for tread depth and miles.

The mean and standard deviation only provide an overall perspective on the data. We also need to attempt to determine if there are any erroneous individual values for our three variables. We start by finding the minimum and maximum values for each variable in Excel. Return to the file *treadwear* and complete the following steps:

Step 1. Enter the heading *Minimum* in cell G4
Step 2. Enter the heading *Maximum* in cell G5
Step 3. Enter =*MIN(C2:C457)* in cell H4
Enter =*MIN(D2:D457)* in cell I4
Enter =*MIN(E2:E457)* in cell J4
Step 4. Enter =*MAX(C2:C457)* in cell H5
Enter =*MAX(D2:D457)* in cell I5
Enter =*MAX(E2:E457)* in cell J5

The results are shown in Figure 4.20. Cells H4 and H5 show that the minimum and maximum values for Life of Tires (Months) are 1.8 months and 601.0, respectively. The minimum value of life of tires in months appears to be reasonable, but the maximum (which is equal to slightly over 50 years) is not a reasonable value for Life of Tires (Months). To identify the automobile with this extreme value, we again sort the entire data set on Life of Tire (Months) and scroll to the last few rows of the data.

We see in Figure 4.20 that the observation with Life of Tire (Months) value of 601.0 is the left rear tire from the automobile with ID Number 8696859. Also note that the left rear tire of the automobile with ID Number 2122934 has a suspiciously high value for Life of Tire (Months) of 111 (almost ten years).

Figure 4.19 Portion of Excel Spreadsheet Showing the Mean and Standard Deviation for Each Variable in the Blakely Tires Data

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)	Tread Depth	Miles
2	80441	LF	19.0	8.1	37419		# of Missing Values	0	0	1
3	80441	LR	19.0	8.1	37419		Mean	23.8	7.6	25440.2
4	80441	RF	19.0	8.1	37419		Standard Deviation	31.83	2.51	23600.21
5	80441	RR	19.0	8.2	37419					
6	95990	LF	8.6	9.7	5670					
7	95990	LR	8.6	9.7	5670					
8	95990	RF	8.6	9.8	5670					
9	95990	RR	8.6	9.7	5670					

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.20 Portion of Excel Spreadsheet for the Blakely Tires Data Sorted by Life of Tire (with rows 8–450 hidden)

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)	Tread Depth	Miles
2	9091771	RF	1.8	10.8	2917		Mean	23.8	7.6	25440.2
3	9091771	RR	1.8	10.7	2917		Standard Deviation	31.83	2.51	23600.21
4	9091771	LF	1.8	10.7	2917		Minimum	1.8	0.0	206.0
5	7712178	LF	2.1	10.7	2186		Maximum	601.0	16.7	107237.0
6	7712178	RR	2.1	10.7	2186					
7	7712178	RF	2.1	10.7	2186					
451	10395904	LF	69.0	0.6	64742					
452	3574739	RR	73.3	0.2	57313					
453	3574739	RF	73.3	0.2	57313					
454	3574739	LF	73.3	0.2	57313					
455	3574739	LR	73.3	0.2	57313					
456	2122934	LR	111.0	9.3	21000					
457	8696859	LR	601.0	2.0	26129					
458										

Note that rows 8-450 are hidden in this figure.

Resorting the data by ID Number and scrolling until we find the four tires from the automobile with ID Number 8696859, we find the value for Life of Tire (Months) for the other three tires from this automobile is 60.1 (see Figure 4.21). This suggests that the decimal for Life of Tire (Months) for this automobile's left rear tire value is in the wrong place.

Figure 4.21 Portion of Excel Spreadsheet for the Blakely Tires Data Sorted by ID Number to Locate records with ID Number 8696859

	A	B	C	D	E
150	8278902	LF	45.7	4.1	11830
151	8278902	LR	45.7	4.1	11830
152	8278902	RF	45.7	4.1	11830
153	8278902	RR	45.7	4.1	11830
154	8696859	LF	60.1	2.0	26129
155	8696859	RF	60.1	2.0	26129
156	8696859	RR	60.1	2.0	26129
157	8696859	LR	601.0	2.0	26129
158	8776539	LF	12.4	9.2	28829
159	8776539	RF	12.4	9.1	28829
160	8776539	RR	12.4	9.1	28829
161	8776539	LR	12.4	9.1	28829
162	8921457	LF	11.7	9.2	19237
163	8921457	LR	11.7	9.3	19237
164	8921457	RR	11.7	9.2	19237
165	8921457	RF	11.7	9.3	19237
166	8952817	LR	52.8	3.0	48961

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Scrolling again to find the four tires from the automobile with ID Number 2122934, we find the value for Life of Tire (Months) for the other three tires from this automobile is 11.1 (see Figure 4.22), which suggests that the decimal for Life of Tire (Months) for this automobile's left rear tire value is also misplaced.

We can now address the extreme values for Life of Tire associated with the Left Rear Tire from the automobile with ID Number 8696859 and the Left Rear Tire from the automobile with ID Number 2122934.

By repeating this process for the remaining variables in the data (Tread Depth and Miles) in columns I and J, we determine that the minimum and maximum values for Tread Depth are 0.0/32nds of an inch and 16.7/32nds of an inch, respectively, and the minimum and maximum values for Miles are 206.0 and 107237.0, respectively (as shown in Figure 4.14). Neither the minimum nor the maximum value for Tread Depth is reasonable; a tire with no tread is not drivable, and the maximum value for tire depth in the data actually exceeds the tread depth on new Blakely brand tires. The minimum value for Miles is reasonable, but the maximum value is not. A similar investigation should be made into these values of tread depth to determine if they are in error and if so, what might be the correct value.

Not all erroneous values in a data set are extreme, and these erroneous values are much more difficult to identify. However, if the variable with suspected erroneous values has a relatively strong relationship with another variable in the data, we can use this knowledge to look for erroneous values. Here we will consider the variables Tread Depth and Miles; because more miles driven should lead to less tread depth on an automobile tire, we expect these two variables to have a negative relationship. A scatter chart, as shown in Figure 4.23, enables us to see whether any of the tires in the data set have values for Tread Depth and Miles that are counter to this expectation.

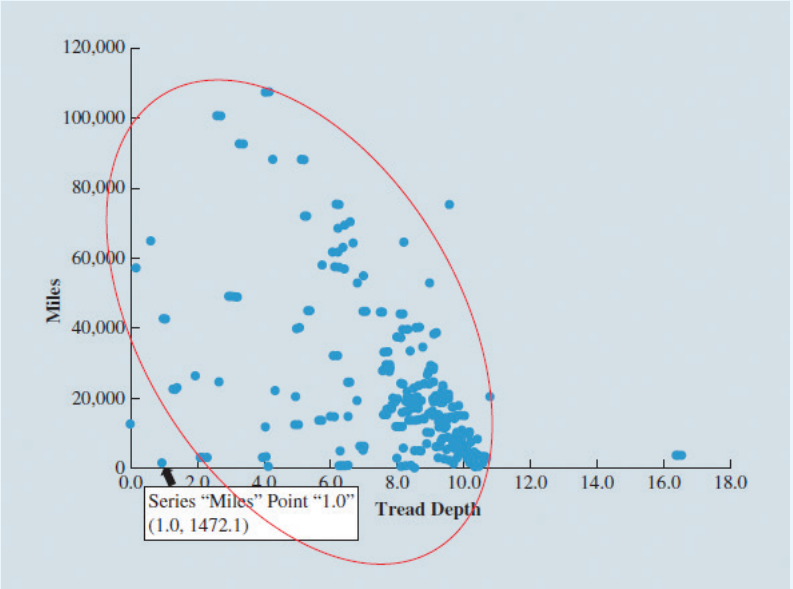
We have drawn a red ellipse in Figure 4.23 around the area in which the records tend to congregate to show the approximate region in which the points representing Tread

Figure 4.22 Portion of Excel Spreadsheet for the Blakely Tires Data Sorted by ID Number to Locate Records with ID Number 2122934

	A	B	C	D	E
42	1383349	RF	25.9	7.1	6094
43	1383349	RR	25.9	7.1	6094
44	1383349	LF	25.9	7.0	6094
45	1383349	LR	25.9	7.1	6094
46	2122934	RF	11.1	9.4	21000
47	2122934	RR	11.1	9.3	21000
48	2122934	LF	11.1	9.3	21000
49	2122934	LR	111.0	9.3	21000
50	2253516	RF	26.2	7.1	55231
51	2253516	LR	26.2	7.1	55231
52	2253516	LF	26.2	7.1	55231
53	2253516	RR	26.2	7.1	55231
54	3121851	LR	17.1	8.4	21378
55	3121851	RR	17.1	8.4	21378
56	3121851	RF	17.1	8.4	21378
57	3121851	LF	17.1	8.5	21378

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.23 Scatter Chart of Tread Depth and Miles for the Blakely Tires Data



Depth and Miles would generally be expected to lie on this scatter plot. The points that lie outside of this ellipse have values for at least one of these variables that is inconsistent with the negative relationship exhibited by the points inside the ellipse. If we position the cursor over one of the points outside the ellipse that is near the origin of the graph, Excel will generate a pop-up box that shows that the values of Tread Depth and Miles for this point are 1.0 and 1472.1, respectively. The tire represented by this point has very little tread and has been driven relatively few miles, which suggests that the value of one or both of these two variables for this tire may be inaccurate and should be investigated.

Closer examination of outliers and potential erroneous values may reveal an error or a need for further investigation to determine whether the observation is relevant to the current analysis. A conservative approach is to create two data sets, one with and one without outliers and potentially erroneous values, and then construct a model on each data set. If a model's implications depend on the inclusion or exclusion of outliers and erroneous values, then you should spend additional time to track down the cause of the outliers.

Once we have resolved the erroneous outlier and other erroneous value issues in the raw data to the best of our ability, we move on to identification of illegitimate duplicate records. In some raw data, duplicate records legitimately occur. But in other raw data, each record is likely unique. Here we demonstrate in Excel how to identify duplicate records with conditional formatting and remove duplicate records with the **Remove Duplicates** tool.

Blakely Tires has now obtained information for each of the four tires from 116 additional automobiles with Blakely brand tires that have been collected through recent state automobile inspection facilities in Texas. The data, contained in the file *treadwearnew*, include the same information as the file *treadwear*. The data are suspicious because they should contain records for 116 sets of four tires or $116 \times 4 = 456$ records, but actually the data include records for 457 tires. We will therefore look for duplicate records in the data.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

We begin by searching for duplicate records using conditional formatting in Excel. After opening the file *treadwearnew:*

DATA*file*
treadwearnew

Step 1. Enter *Duplicate Record?* In cell F1
Step 2. Enter *=IF(COUNTIFS($A$2:$A$458,$A2,$B$2:$B$458,$B2,$C$2:$C$458, $C2,$D$2:$D$458,$D2,$E$2:$E$458,$E2)>1, "Duplicate record", "")* in cell F2

The IF function is discussed in more detail in Chapter 12.

The COUNTIFS function applies criteria to cells across multiple ranges and counts the number of times all criteria are met. By nesting the COUNTIFS function inside the IF function in step 2, we will determine if each record in the range A2:A458 is a duplicate on the criteria in ranges B2:B458, C2:C458, D2:D458, and E2:E458 and we enter *Duplicate record* in the corresponding cell in column F if the corresponding record is a duplicate on all of these four ranges.

Step 3. Copy cell F2 to cells F3:F458

Each cell in the range F2:F458 that corresponds to a range that contains a duplicate record will now contain the phrase Duplicate record, and all other cells in this range will remain blank.

Step 4. Select cells A2:F458
Step 5. Click the **Data** tab on the ribbon
Step 6. Click the **Sort** button in the **Sort & Filter** group
Step 7. In the **Sort** dialog box
Select **Duplicate Record?** in the **Sort by** window
Select **Z to A** in the **Order** window
Click the **Add Level** button
Select **ID Number** in the new **Sort by** window
Select **Smallest to Largest** in the **new Order** window
Click **OK**

As Figure 4.24 shows, the duplicate records, sorted by ID Number, appear in the first rows of the data in the worksheet.

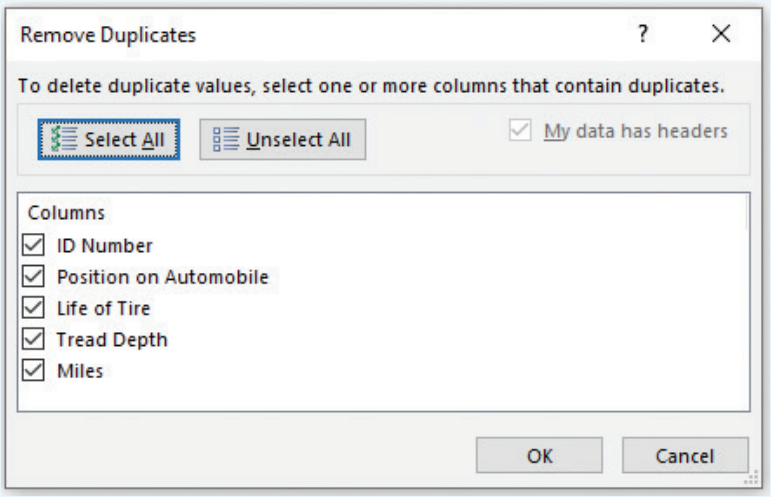
We can now see that the vehicle with ID Number 81518 has duplicate records for its right rear tire and unique records for each of the other three positions on the automobile. We also see that these data do not have any other duplicate records. We can now delete one of the two records for the right rear tire of the vehicle with ID Number 81518 and save the file.

Figure 4.24 Identification of Duplicate Records in the New Blakely Tires Data

	A	B	C	D	E	F
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles	Duplicate Record?
2	81518	RR	20.6	8.6	37408	Duplicate record
3	81518	RR	20.6	8.6	37408	Duplicate record
4	13391952	LR	58.5	2.2	3007	
5	21679618	LR	17.2	8.0	39084	
6	18414342	RR	15.0	9.8	13709	
7	19778015	RR	7.3	10.9	2002	
8	16356967	RR	13.3	8.5	24334	
9	8951784	LR	53.0	3.2	49033	
10	6558415	RR	14.0	8.4	4296	
11	16289891	LR	5.4	10.4	5647	
12	8698313	LF	61.1	1.6	26190	
13	8919904	LF	11.4	10.1	19490	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.25 Remove Duplicates Dialog Box



Once they are identified, manual deletion of duplicate records is simple. But if there are many duplicate records, a more automated approach will be more efficient. In Excel, we can accomplish this task by using the **Remove Duplicates** tool. With the file *treadwearnew* open:

Step 1. Select cells A2:E458
Step 2. Click the **Data** tab on the ribbon
Step 3. In the **Data Tools** group, click the **Remove Duplicates** button
Step 4. When the **Remove Duplicates** dialog box (Figure 4.25) opens:
Click the **Select All** button to indicate you want to find rows that are duplicates on all columns
Click **OK**

This action will delete all but one record from each set of duplicate records. This deletes one of the two duplicate rows for the vehicle with ID Number 81518.

Note that although you can skip the steps to detect duplicate records and proceed directly to deletion of duplicate records, this is a risky strategy. You should always identify and examine duplicate records to determine whether they are legitimately separate records before deleting them from the data.

Notes + Comments

1. Many of the other descriptive statistics tools described in Chapter 2 and data visualization tools described in Chapter 3 can be used to aid in data cleaning.
2. Many software packages used for analyzing data, such as R and Python, can be used to clean data. The exact steps for doing this depend on the software tool that is used.

4.4 Enriching

The discovery, structuring, and cleaning steps of the data wrangling process provide a deeper understanding of the raw data. In the enriching step, we use this understanding in thinking about what information would be useful to have in the data set. Here we consider

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

how we might want to simplify the data by selecting the subset of fields or records that will be useful for the ensuing analyses, supplement the raw data with additional data, or enhance the raw data through transformations on its fields. Although enriching the data is an optional step, it can be an extremely important step if the raw data does not meet requirements for the ensuing analysis.

Subsetting Data

Subsetting is the extraction of fields and records that will be useful for ensuing analyses from data. This creates a more manageable data set that can be processed and analyzed more efficiently. Some software subsets data by selecting or retaining the fields and records that will be useful for ensuing analyses, and other software subsets data by excluding or dropping the fields and records that will not be useful for ensuing analyses. In either approach, the result is a new data set that contains only the fields and records that are to be used in the upcoming analyses.

Suppose we want to consider only tires that were in the left front position and had tread depth of no more than 2/32nds of an inch. We could use the following steps to produce the subset of records that satisfy these conditions. After opening the file *treadwear*:

DATA *file*
treadwear

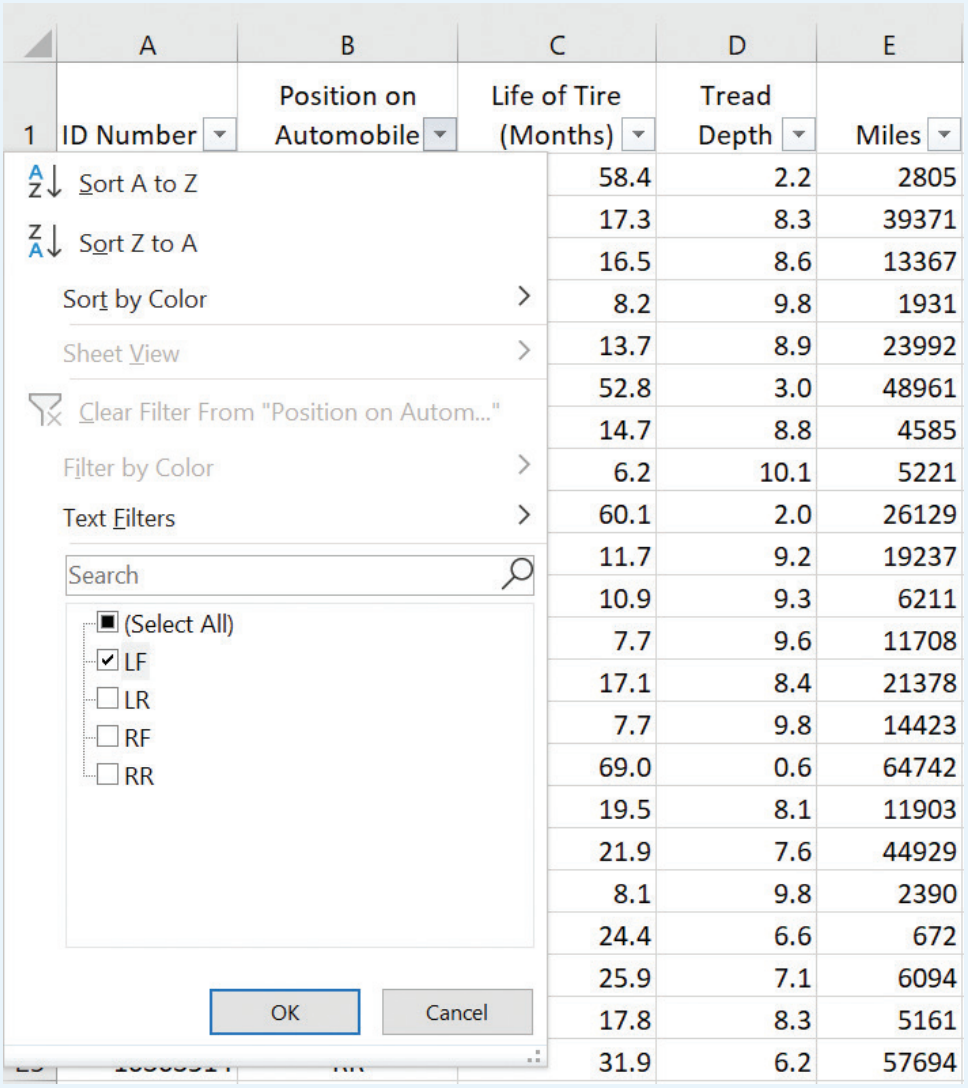
Step 1. Select cells A1:E458
Step 2. Click the **Data** tab on the Ribbon
Step 3. In the **Sort & Filter** group click the **Filter** button to enable the filters (these can be used to sort and filter the records of the data on values of their respective columns) as shown in Figure 4.26
Step 4. Click the drop-down arrow in cell B1, select **Test Filters**, then select **Equals…** to open the **Custom AutoFilter** dialog box
Step 5. In the **Custom AutoFilter** dialog box (Figure 4.27)
In the top box under **Show rows where: Position on Automobile**, click the down arrow and select equals
In the adjacent box**,** click the down arrow and select LF
Click **OK**

Figure 4.26 The Blakely Tires Data with Filters Enabled

	A	B	C	D	E
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles
2	13391487	LR	58.4	2.2	2805
3	21678308	LR	17.3	8.3	39371
4	18414311	RR	16.5	8.6	13367
5	19778103	RR	8.2	9.8	1931
6	16355454	RR	13.7	8.9	23992
7	8952817	LR	52.8	3.0	48961
8	6559652	RR	14.7	8.8	4585
9	16289814	LR	6.2	10.1	5221

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.27 The Custom AutoFilter Dialog Box for Position on Automobile



Step 6. Click the drop-down arrow in cell D1, select **Number Filters**, then select **Less than or equal to** open the **Custom AutoFilter** dialog box

Step 7. In the **Custom AutoFilter** dialog box

In the top box under **Show rows where: Tread Depth**, click the down arrow and select **is less than or equal to**

In the adjacent box**,** enter *2*

Click **OK**

The resulting display as shown in Figure 4.28, includes only records that correspond to tires that were on the left front position and had tread depth of no more than 2/32nds of an inch.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.28 The Blakely Tires Data Filtered to Include Only Left Front Tires with Tread Depth of No More than 2/32nds of an Inch

	A	B	C	D	E
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles
10	8696859	LF	60.1	2.0	26129
70	14922617	LF	39.6	0.0	12346
159	16691383	LF	66.5	1.1	42632
294	10395904	LF	69.0	0.6	64742
401	3574739	LF	73.3	0.2	57313
436	21817569	LF	64.1	1.3	22618
458					

Step 8. In the **Custom AutoFilter** dialog box
In the top box under **Show rows where: Tread Depth**, click the down arrow and select **is less than or equal to**
In the adjacent box**,** enter *2*
Click **OK**

To restore all data to our view, click the **Home** *tab on the Ribbon, then click the Filter button in the* **Sort & Filter** *group to disable the filters on the worksheet.*

Note that all 458 records are still present on this worksheet, but all are hidden except those that satisfy the filters we applied. To perform analyses on only the filtered data in Excel, we must select the cells with the filtered data, copy the filtered data and paste then into a new worksheet, then perform the analyses on the data in the new worksheet.

Subsetting can also be used to support analyses of groups of interest in a data set. For example, we could calculate the mean life of tire for the filtered data:

$$\bar{x} = \frac{372.6}{6} = 62.1$$

This is the sample mean life of tires data that were in the left front position and had tread depth of no more than 2/32nds of an inch. We could then follow the same process to produce the sample mean for records that correspond to tires that were on the right front position and had tread depth of no more w/32nds of an inch to assess the difference in mileage for tires that were on the two front positions and had tread depth of no more than 2/32nds of an inch.

Supplementing Data

We often identify additional data that we want to add to the records in the raw data or receive raw data in multiple files. In these instances, these data must be combined into a single data set if they are all potentially relevant to the ensuing analyses.

If we have different data for individual records in multiple files, we combine the data from these files by appending fields. If each record occurs no more than once in each data set, we combine the data for each unique record by using a **one-to-one merger** of the data sets. If we have different data for individual records in multiple locations and at least one record occurs more than once in at least one of the data sets, we combine the data for each unique record by using a **one-to-many merger** of the data sets. These operations are generally similar.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The XLOOKUP function is discussed in more detail in Chapter 12.

The file *treadwearmodels* contains information on the model for each vehicle in the *treadwear* data by ID Number as shown in Figure 4.29. Here we want to append the appropriate value for the field Model in the file *treadwearmodels* to each record in the *treadwear* data. In the following steps, we demonstrate how to do this using Excel's XLOOKUP function. This is an example of a one-to-many merger since four records are associated with each ID Number.

After opening the files *treadwear* and *treadwearmodels*

DATA*file*
treadwear

DATA*file*
treadwearmodels

Step 1. Enter *Model* in cell F1 of the DATA worksheet in the file *treadwear*

Step 2. Enter *=XLOOKUP(A2,[treadwearmodels.xlsx]Sheet4!$A$2:$A$115, [treadwearmodels.xlsx]Sheet4!$B$2:$B$115)* in cell F2 of the *treadwear* spreadsheet where

- A2 is the cell that contains the value to be matched
- [treadwearmodels.xlsx]Sheet4!$A$2:$A$115 identifies the cells that will be searched for a matching value
- [treadwearmodels.xlsx]Sheet4!$B$2:$B$115 identifies the cells that contain the values to be transferred if a match is found

Step 3. Copy cell F2 into cells F3:F457 of the *treadwear* spreadsheet

In earlier versions of Excel, VLOOKUP allowed the user to search a data range that is set up with fields in columns, and HLOOKUP allowed the user to search a data range that is set up with fields in rows. XLOOKUP is capable of searching a data range that is set up with the fields in columns or rows.

Figure 4.30 shows partial results of the one-to-many merger of the *treadwearmodels* data with the *treadwear* data. Note that the XLOOKUP function can merge data from multiple worksheets or from multiple workbooks.

We sometimes have multiple data sets that each contain unique records. This often happens with time series data. For example, we may have three separate files that each contain quarterly sales data for one of the past three years as shown in Figure 4.31.

In Excel, we deal with cases such as this by cutting and pasting the raw data from these files into a new file in which each record represents a unique record from one of the ten original raw data files. The result of combining the three years of quarterly sales data in Figure 4.31 into a single worksheet is shown in Figure 4.32.

DATA*file*
qtrsalesbyear

Figure 4.29 A Portion of the Blakely Tires Model Data

	A	B
1	ID Number	Model
2	80441	Volkswagen
3	95990	Chevrolet
4	202372	Cadilac
5	238850	Dodge
6	292849	Kia
7	662056	Chevrolet
8	852812	Chevrolet
9	888432	Toyta
10	1064251	Chevrolet
11	1086054	Ford
12	1383349	Mercedes
13	2122934	Dodge

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.30 A Portion of the Blakely Tires Data with the Field Model Appended to Each Record

	A	B	C	D	E	F
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles	Model
2	13391487	LR	58.4	2.2	2805	Hyundai
3	21678308	LR	17.3	8.3	39371	Mercedes
4	18414311	RR	16.5	8.6	13367	Volkswagen
5	19778103	RR	8.2	9.8	1931	Volvo
6	16355454	RR	13.7	8.9	23992	Volkswagen
7	8952817	LR	52.8	3.0	48961	Chrysler
8	6559652	RR	14.7	8.8	4585	Chevrolet
9	16289814	LR	6.2	10.1	5221	Acura
10	8696859	LF	60.1	2.0	26129	Volkswagen
11	8921457	LF	11.7	9.2	19237	Chevrolet
12	7505467	LF	10.9	9.3	6211	Chevrolet
13	18307885	RR	7.7	9.6	11708	Acura

Figure 4.31 Three Years of Quarterly Sales Data in Separate Files

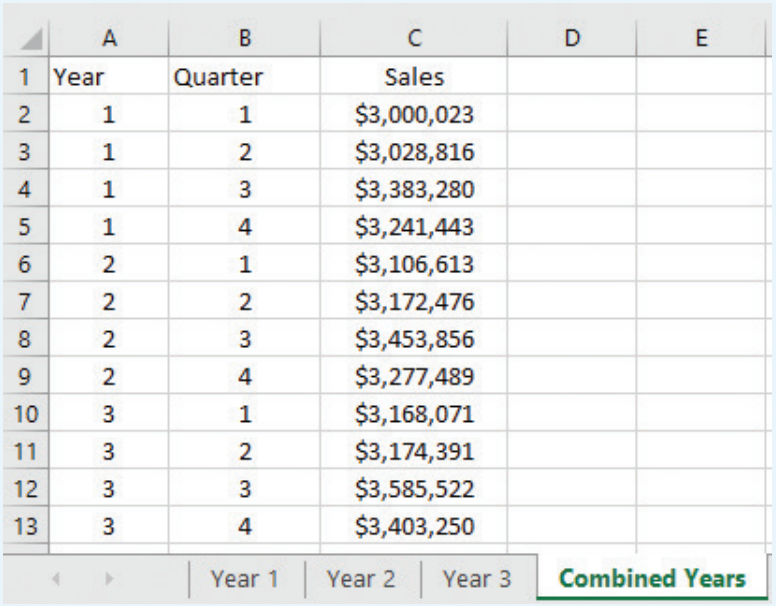
Year	Quarter	Sales
1	1	$3,000,023
1	2	$3,028,816
1	3	$3,383,280
1	4	$3,241,443

Year	Month	Sales
2	1	$3,106,613
2	2	$3,172,476
2	3	$3,453,856
2	4	$3,277,489

Year	Month	Sales
3	1	$3,168,071
3	2	$3,174,391
3	3	$3,585,522
3	4	$3,403,250

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.32 Three Years of Quarterly Sales Data in a Single Worksheet



	A	B	C	D	E
1	Year	Quarter	Sales		
2	1	1	$3,000,023		
3	1	2	$3,028,816		
4	1	3	$3,383,280		
5	1	4	$3,241,443		
6	2	1	$3,106,613		
7	2	2	$3,172,476		
8	2	3	$3,453,856		
9	2	4	$3,277,489		
10	3	1	$3,168,071		
11	3	2	$3,174,391		
12	3	3	$3,585,522		
13	3	4	$3,403,250		

Year 1 | Year 2 | Year 3 | **Combined Years**

Other software packages can deal with this issue in a more direct and less manual manner.

Enhancing Data

When we make a transformation on one or more fields in a data set, we change the format, structure, or values of data in a well-defined systematic manner. This is sometimes done for presentation; for example, an analyst may reformat a field to show as a percentage because that will communicate more effectively with the target audience. However, transformations on fields are also frequently used to facilitate modeling. In this section, we will discuss transforming quantitative fields and transforming categorical fields.

In Chapter 2, we defined quantitative data as data for which numerical values are used to indicate magnitude, such as how many or how much. Recall that arithmetic operations such as addition, subtraction, and multiplication can be performed on quantitative data. This allows us to create new quantitative fields as functions of existing quantitative fields using these mathematical operations. Most data wrangling software packages provide tools for making functional transformations of quantitative fields.

Data sets often contain quantitative fields that, considered separately, are not particularly insightful but that when appropriately combined, result in a new field that may be useful in the ensuing analyses. For example, suppose an analyst wanted to use the miles per months achieved by each tire in the Blakely tires data shown in Figure 4.30. The analyst could use the following steps in Excel to do so. With the file *treadwear* open:

Step 1. Enter *Miles per Month* in cell G1 of the DATA worksheet
Step 2. Enter $=E2/C2$ in cell F2
Step 3. Copy cell F2 into cells F3:F457

Figure 4.33 shows a portion of the Blakely Tires data with the Miles per Month Field added.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.33 A Portion of the Blakely Tires Data with the Miles per Month Field Added

	A	B	C	D	E	F	G
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles	Model	Miles per Month
2	13391487	LR	58.4	2.2	2805	Hyundai	48.03
3	21678308	LR	17.3	8.3	39371	Mercedes	2275.78
4	18414311	RR	16.5	8.6	13367	Volkswagen	810.12
5	19778103	RR	8.2	9.8	1931	Volvo	235.49
6	16355454	RR	13.7	8.9	23992	Volkswagen	1751.24
7	8952817	LR	52.8	3.0	48961	Chrysler	927.29
8	6559652	RR	14.7	8.8	4585	Chevrolet	311.90
9	16289814	LR	6.2	10.1	5221	Acura	842.10
10	8696859	LF	60.1	2.0	26129	Volkswagen	434.76
11	8921457	LF	11.7	9.2	19237	Chevrolet	1644.19
12	7505467	LF	10.9	9.3	6211	Chevrolet	569.82
13	18307885	RR	7.7	9.6	11708	Acura	1520.52

In other instances, a functional transformation of an existing field may result in a new field that reveals an important relationship that may be useful in the ensuing analyses. We can apply a wide variety of functional transformations, such as a logarithmic transformation, to an existing quantitative field to create a new quantitative field that may be relevant to the analysis of the data. Here we demonstrate how to apply a natural logarithmic transformation to the field Tread Depth. With the file *treadwear* open:

The Excel function LN calculates the natural logarithm of a number.

Step 1. Enter *LN Tread Depth* in cell H1 of the DATA worksheet
Step 2. Select cell H2
Step 3. Click the **Formulas** tab and click the **Insert Function** button f_x to open the **Insert Function** dialog box
In the **Or select a category** dropdown menu, select **Math & Trig**
In the **Select a function** area, scroll down to **LN**
Select **LN**
Click **OK**
Step 4. When the **Function arguments** dialog box appears, enter *D2* into the **Number** box
Click **OK**
Step 5. Copy cell H2 into cells H3:H457

Figure 4.34 shows a portion of the *treadwear* data with the LN Tread Depth field added.

Next we will create supplementary data for a field that is a categorical variable. In Chapter 2, we defined categorical data as data for which categories of like items are identified by labels or names. Recall that arithmetic operations cannot be performed on categorical data (even if numbers are used as the labels).

Dummy variables are used extensively for analytics including regression models and data mining applications.

It is sometimes necessary to create a **dummy variable**, or a field that takes values of 0 or 1 to indicate the absence or presence of some categorical effect, from the values of an

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.34 A Portion of the Blakely Tires Data with the LN Tread Depth Field Added

	A	B	C	D	E	F	G	H
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles	Model	Miles per Month	LN Tread Depth
2	13391487	LR	58.4	2.2	2805	Hyundai	48.03	0.7884574
3	21678308	LR	17.3	8.3	39371	Mercedes	2275.78	2.1162555
4	18414311	RR	16.5	8.6	13367	Volkswagen	810.12	2.1517622
5	19778103	RR	8.2	9.8	1931	Volvo	235.49	2.2823824
6	16355454	RR	13.7	8.9	23992	Volkswagen	1751.24	2.1860513
7	8952817	LR	52.8	3.0	48961	Chrysler	927.29	1.0986123
8	6559652	RR	14.7	8.8	4585	Chevrolet	311.90	2.1747517
9	16289814	LR	6.2	10.1	5221	Acura	842.10	2.3125354
10	8696859	LF	60.1	2.0	26129	Volkswagen	434.76	0.6931472
11	8921457	LF	11.7	9.2	19237	Chevrolet	1644.19	2.2192035
12	7505467	LF	10.9	9.3	6211	Chevrolet	569.82	2.2300144
13	18307885	RR	7.7	9.6	11708	Acura	1520.52	2.2617631
14	3121851	LR	17.1	8.4	21378	Kia	1250.18	2.1282317

existing field in the raw data. This is usually done by defining a new field as a function of the values of an existing field in the raw data. We proceed with a few examples from Excel for the data in the file *airportconditions*, which contains daily information on several weather conditions for a regional airport in the Midwest for February. These data are shown in Figure 4.35.

DATA*file* airportconditions

We will first create a dummy variable for a quantitative variable; this dummy variable will indicate whether the temperature dropped below freezing during each day. The dummy variable, which we will name Below32, will equal 1 if the low temperature at the airport was below 32 during each date and 0 otherwise. Once the file *airportconditions* is open in Excel:

Step 1. Enter *Below32* In cell F1
Step 2. Enter *=IF(C2<32,1,0)* in cell F2
Step 3. Copy cell F2 into cells F3:F29

We have created our dummy variable Below32 in column F.

The IF *function is discussed in more detail in Chapter 12.*

We will next create a series of separate dummy variables for a categorical variable that has multiple values. These dummy variables will indicate whether runway conditions were rainy, icy, or dry for each date.

Step 1. Enter *Rainy* in cell G1, *Icy* in cell H1, and *Dry* in cell I1
Step 2. Enter *=IF(D2="Rainy",1,0)* in cell G2
Step 3. Enter *=IF(D2="Icy",1,0)* in cell H2
Step 4. Enter *=IF(D2="Dry",1,0)* in cell I2
Step 5. Copy cells G2:I2 into cells G3:I29

We have created our dummy variables Rainy, Icy, and Dry in columns G, H, and I that are equal to 1 if the associated condition (Rainy, Icy, Dry) was present on a particular date and 0 otherwise.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.35 Weather Conditions for a Regional Airport in the Midwest During February of Last Year

	A	B	C	D	E
1	Date	High Temperature	Low Temperature	Conditions	Wind Speed (mph)
2	2/1	43	21	Rainy	21
3	2/2	66	48	Rainy	27
4	2/3	58	43	Rainy	15
5	2/4	49	8	Dry	15
6	2/5	65	45	Dry	12
7	2/6	42	21	Dry	10
8	2/7	48	30	Icy	23
9	2/8	42	40	Rainy	6
10	2/9	46	42	Dry	5
11	2/10	37	31	Dry	13
12	2/11	20	16	Icy	15
13	2/12	33	0	Dry	24
14	2/13	44	22	Rainy	5
15	2/14	38	24	Dry	0
16	2/15	50	40	Rainy	5
17	2/16	53	46	Rainy	10
18	2/17	46	38	Rainy	4
19	2/18	45	29	Dry	15
20	2/19	42	29	Icy	26
21	2/20	35	11	Icy	19
22	2/21	34	20	Dry	15
23	2/22	44	16	Dry	1
24	2/23	34	16	Dry	23
25	2/24	39	38	Rainy	8
26	2/25	50	41	Dry	14
27	2/26	51	47	Dry	7
28	2/27	49	30	Rainy	18
29	2/28	25	5	Dry	15

Next, we will create a dummy variable based on the values of two variables, one quantitative and one categorical. If runway conditions are icy and the windspeed is at least 20 MPH, hazardous landing conditions exist. The dummy variable we will create here indicates whether runway conditions were rainy and there was high wind for each date.

Step 1. Enter *Hazardous Conditions* in cell J1
Step 2. Enter *=IF(D2="Icy",IF(E2>=20,1,0),0)* in cell J2
Step 3. Copy cell J2 into cells J3:J29

We have created our dummy variable in column J that is equal to 1 if hazardous landing conditions exist on a particular date and 0 otherwise

Finally, we will create a categorical variable that represents a reduction in the number of categories for the field Conditions. The airport communications system automatically sends a text message to the crew of an outbound flight if the airport is experiencing precipitation (Rainy or Icy) on the date of the flight. The categorical variable we will create

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 4.36 Weather Conditions Data with Dummy Variables

	A	B	C	D	E	F	G	H	I	J	K
1	Date	High Temperature	Low Temperature	Conditions	Wind Speed	Below32	Rainy	Icy	Dry	Hazardous Conditions	Precipitation?
2	2/1	43	21	Rainy	21	1	1	0	0	0	Precipitation
3	2/2	66	48	Rainy	27	0	1	0	0	0	Precipitation
4	2/3	58	43	Rainy	15	0	1	0	0	0	Precipitation
5	2/4	49	8	Dry	15	1	0	0	1	0	No Precipitation
6	2/5	65	45	Dry	12	0	0	0	1	0	No Precipitation
7	2/6	42	21	Dry	10	1	0	0	1	0	No Precipitation
8	2/7	48	30	Icy	23	1	0	1	0	1	Precipitation
9	2/8	42	40	Rainy	6	0	1	0	0	0	Precipitation
10	2/9	46	42	Dry	5	0	0	0	1	0	No Precipitation
11	2/10	37	31	Dry	13	1	0	0	1	0	No Precipitation
12	2/11	20	16	Icy	15	1	0	1	0	0	Precipitation
13	2/12	33	0	Dry	24	1	0	0	1	0	No Precipitation
14	2/13	44	22	Rainy	5	1	1	0	0	0	Precipitation
15	2/14	38	24	Dry	0	1	0	0	1	0	No Precipitation
16	2/15	50	40	Rainy	5	0	1	0	0	0	Precipitation
17	2/16	53	46	Rainy	10	0	1	0	0	0	Precipitation
18	2/17	46	38	Rainy	4	0	1	0	0	0	Precipitation
19	2/18	45	29	Dry	15	1	0	0	1	0	No Precipitation
20	2/19	42	29	Icy	26	1	0	1	0	1	Precipitation
21	2/20	35	11	Icy	19	1	0	1	0	0	Precipitation
22	2/21	34	20	Dry	15	1	0	0	1	0	No Precipitation
23	2/22	44	16	Dry	1	1	0	0	1	0	No Precipitation
24	2/23	34	16	Dry	23	1	0	0	1	0	No Precipitation
25	2/24	39	38	Rainy	8	0	1	0	0	0	Precipitation
26	2/25	50	41	Dry	14	0	0	0	1	0	No Precipitation
27	2/26	51	47	Dry	7	0	0	0	1	0	No Precipitation
28	2/27	49	30	Rainy	18	1	1	0	0	0	Precipitation
29	2/28	25	5	Dry	15	1	0	0	1	0	No Precipitation

here indicates whether runway conditions were rainy or icy (Precipitation) or dry (No Precipitation) for each date.

Step 1. Enter *Precipitation?* in cell K1
Step 2. Enter *=IF(D2="Icy","Precipitation",IF(D2="Rainy","Precipitation","No Precipitation"),"No Precipitation")* in cell K2
Step 3. Copy cell K2 into cells K3:K29

We have created our categorical variable in column K that is equal to Precipitation if conditions are rainy or icy on a particular date and No Precipitation otherwise. Figure 4.36 shows the resulting Excel worksheet with the additional dummy variables and categorical variable.

4.5 Validating and Publishing

We conclude the steps in the data wrangling process with ensuring the data are accurate and accessible. This is accomplished through validating and publishing.

Validating

In the validation step, the analyst checks the accuracy and consistency of data before it is made available for analysis or used for other business operations. This includes i) checking that the information in the wrangled data correctly represents the information provided by the raw data and ii) determining that the fields in the data sets (particularly new fields created in the data wrangling process) are defined in the same manner for all records in the data.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Many data wrangling software packages allow for the identification of values of a field that are outside of the known minimum and maximum values for that field. If you know the minimum and maximum values of a field, these tools can allow you to identify such invalid values in a more direct and efficient manner. In Excel, this is done using the Data Validation tool. We demonstrate the use of the Excel Data Validation tool using the Tread Depth field in the Blakely Tire data. We know that no tire can have a nonpositive value or a value greater than 10 (the tread depth of new Blakely brand tires). Once the file *treadwear* is open, complete the following steps:

DATA*file*
treadwear

Step 1. Select cells C2:C457

Step 2. Click the **Data** tab on the Ribbon

Step 3. In the **Data Tools** group, click the **Down Arrow** on the Data **Validation** button and select **Data Validation** (note that clicking the **Data Validation** button rather than its **Down Arrow** will open the **Data Validation** dialog box directly)

Step 4. When the **Data Validation** dialog box opens:
Select **Decimal** in the **Allow** box
Select **Between** in the **Data** box
Enter *0.1* in the **Minimum** box
Enter *10* in the **Maximum** box
Click **OK**

Step 5. In the **Data Tools** group, click the **Down Arrow** on the **Validation** button and select **Circle Invalid Data** (do not click the **Validation** button rather than its **Down Arrow**, as this will open the **Data Validation** dialog box directly)

We immediately note that the cell showing Tread Depth for the left rear tire with ID Number 16289814 is circled as shown in Figure 4.37. Scrolling through the data reveals several other tires with Tread Depth values outside of the valid range we specified for the field.

Figure 4.37 Portion of Excel Spreadsheet for the Blakely Tire Data Showing an Extreme Value of Tread Depth Circled

	A	B	C	D	E
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles
2	13391487	LR	58.4	2.2	2805
3	21678308	LR	17.3	8.3	39371
4	18414311	RR	16.5	8.6	13367
5	19778103	RR	8.2	9.8	1931
6	16355454	RR	13.7	8.9	23992
7	8952817	LR	52.8	3.0	48961
8	6559652	RR	14.7	8.8	4585
9	16289814	LR	6.2	10.1	5221
10	8696859	LF	60.1	2.0	26129
11	8921457	LF	11.7	9.2	19237
12	7505467	LF	10.9	9.3	6211
13	18307885	RR	7.7	9.6	11708

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note that the circles created by Excel's **Data Validation** tool are static; if you sort the data, the circles will remain in their original locations and will no longer be meaningful. Thus, if you want to use this tool to identify extreme values on data sorted in a particular order, you must sort the data so they are in the desired order prior to applying the Excel's **Data Validation** tool. To remove the circles, click the **Down Arrow** on the **Data Validation** button in the **Data Tools** group and select **Clear Validation Circles** (do not click the **Data Validation** button rather than its **Down Arrow**, as this will open the **Data Validation** dialog box directly).

The validation step is similar to the cleaning step, and we can use the same tools we used for cleaning and validating the data (summary statistics, frequency distributions, bar charts and histograms, *z*-scores, scatter charts, correlation coefficients, and Excel's Data Validation and Remove Duplicates tools).

As with the cleaning step, the validation step may need to be repeated several times to ensure the accuracy and consistency of the data. Validation is the final step in preparing the raw data for analysis, so this is the last opportunity for the analyst to identify and address issues with the data that are to be analyzed.

It is often the case that anomalies in the data persist after the initial wrangling of the data. These anomalies might only be discovered once the analysis is under way and unexpected results occur. In these cases, the data wrangling and analysis might take several iterations.

Publishing

Once the raw data have been wrangled into a high-quality data set that is ready for analyses, we enter the publishing step of data wrangling. At this point, we make the newly wrangled data available in a location that make it easy to access and in a format that is easy to use. If we fail to accomplish either of these goals in the publishing step, all the work we performed in the first five steps of the data wrangling process is fruitless.

Software packages used for data wrangling are generally capable of exporting wrangled data in a wide variety of formats, including:

- Delimited text (txt) files, comma separated value (CSV) files, JavaScript Object Notation (JSON) files, XML files, Excel (xlsx) files, and HTML files
- Databases such as ACCESS, SQL, or ORACLE
- Web pages/HTML files

In addition to making the wrangled data available in a format suitable for analyses, a data dictionary must be created for the data in the publishing step. The **data dictionary** documents characteristics of the data such as names and definitions of the fields, units of measure used in the fields, the source(s) of the raw data, relationship(s) of the wrangled data with other data, and other attributes. Thus, an effective data dictionary is critical because it ensures that everyone who uses the wrangled data can understand the data and use them in appropriate ways across various analyses.

Summary

In this chapter, we have introduced data wrangling. We discussed the importance of data wrangling to the quality and meaningfulness of analyses to be performed using the data, and the consequences of devoting insufficient time and effort into wrangling the data. We then proceeded to a detailed discussion of the six steps in data wrangling, beginning with the discovery step and consideration of accessing the raw data and the format of the raw data. We then proceeded to the structuring step of the data wrangling process and consideration of formatting the data, the arrangement of data, splitting a single field into multiple fields, and combining multiple fields into a single field.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In our discussion of the cleaning step, we dealt with finding and dealing with missing data including methods for data imputation. In this step, we also discussed identification of erroneous values, outliers, and duplicate records. We considered subsetting, supplementing, and enhancing the raw data in our discussion of the enriching step of the data wrangling process. Here we also considered creating new quantitative fields and new categorical fields. We concluded with a discussion of the validating and publishing steps of the data wrangling process.

Glossary

Cleaning The data wrangling step in which errors in the raw data are corrected.
Data dictionary Documentation of characteristics of the wrangled data such as names and definitions of the fields, units of measure used in the fields, the source(s) of the raw data, relationship(s) of the wrangled data with other data, and other attributes.
Data wrangling The process of cleaning, transforming, and managing data so it is more reliable and can be more easily accessed and used for analysis.
Delimiter A tag or marker that separates structured data into various fields.
Discovery The data wrangling step in which the analyst becomes familiar with the data in order to conceptualize how it might be used and potentially discovers issues that will need to be addressed later in the data wrangling process.
Dummy variable A field that that takes a value of 0 or 1 to indicate the absence or presence of some categorical effect.
Enriching The data wrangling step in which the raw data are augmented by incorporating values from other data sets and/or applying transformations to portions of the existing data to ensure that all data that will be required for the ensuing analyses will be included in the resulting data set.
Field A characteristic of the observations in a data set.
Flat file A data file in which structured data are arrayed as a rectangle, with each row representing an observation or record, and each column representing a unique variable or field.
Illegitimately missing data Instances for which there is an appropriate reason for the value of a field to be missing.
Imputation Systematic replacement of missing values with values that seem reasonable.
Legitimately missing data Instances for which there is an appropriate reason for the value of a field to be missing.
Missing at random (MAR) Instances for which the tendency for a record to be missing a value of some field is related to the value of some other fields(s) in the record.
Missing completely at random (MCAR) Instances for which the tendency for a record to be missing a value of some field is entirely random.
Missing not at random (MNAR) Instances for which the tendency for a record to be missing a value of some field is related to the missing value.
Numeric data are data that are stored in a manner that allows mathematical operations to be performed on them. Data of this type generally represent a count or measurement.
One-to-one merger Combining multiple data sets that each have different data for individual records, when each record occurs no more than once in each data set.
One-to-many merger Combining multiple data sets that each have different data for individual records, when at least one record occurs more than once in at least one of the data sets
Publishing The data wrangling step in which a file containing the wrangled data and documentation of the file's contents are made available to its intended users in a format they can use.
Raw data Data that has not been processed or prepared for analysis.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Record A grouping of characteristics for a particular observation in a data set.
Semi-structured data Data that does not have the same level of organization as structured data, but that allow for isolation of some elements of the raw data when they are imported.
Stacked data Data organized so that the values for each variable are stored in a single field.
Structured data Data sets that are arrayed in a predetermined pattern that make them easy to manage and search.
Structuring The data wrangling step in which the raw data file is arranged so it can be more readily analyzed in the intended manner.
Subsetting The extraction of fields and records that will be useful for ensuing analyses from data.
Text data Data that are words, phrases, sentences, and paragraphs.
Unstacked data Data organized so that the values for one variable correspond to separate fields, each of which contains observations corresponding to these respective fields.
Unstructured data Databases that are not arranged in a predetermined pattern, and therefore in their raw form cannot be stored in a manner similar to a flat file.
Validating The data wrangling step in which the accuracy and reliability of the wrangled data are confirmed so they are ready for the ensuing analyses.

Problems

1. **Six Steps of Data Wrangling.** List the steps of data wrangling and explain what is done in each step. LO 1

DATA*file*
detprod.csv, albprod.csv, proprod.csv

2. **Combining Monthly Production Reports.** Each month, Morgan Fabrication receives csv files from its facilities in Detroit, Albany, and Providence. These files summarize production in each facility for each day during the previous month, and each of these files contains the same fields (date and production of each of five fabricated parts for each day of the month) and are arranged in the same format. Information on January production for this year at these three facilities is provided in the files *detprod.csv*, *albprod.csv*, and *proprod.csv*. Import the data from these files, combine them to create a single set of data, and save the result as *janprod.xlsx*. LO 2, 7

3. **Combining Lawn Treatment Records.** Lush&Green Lawn Services employs four lawn treatment specialists: Penny Esposito, Garland Falkland, Peter Raft, and Baozhai Suen. Each specialist drives a truck equipped with a small computer they use to record the details of the treatments they apply to each property. At the end of each workday, the computer on each truck uploads the data for the day to a cloud server, where the data are available in tab delimited files named *esposito.txt*, *falkland.txt*, *raft.txt*, and *suen.txt*. Import the data from these files, combine them to create a single set of data, and save the result as *treatments.xlsx*. LO 2, 7

DATA*file*
esposito.txt, falkland.txt, raft.txt, suen.txt

4. **Unstacking Marketing Research Data.** Brugha Research Associates (BRA) just completed a controlled store test of a proposed new packaging for the Aussie brand of cookies. BRA collected eight weeks of pretest period sales data for Aussie cookies from 20 selected grocery stores in the Salt Lake City area, randomly divided these stores into two panel stores of 10 grocery stores, and designated one panel as the test panel and the other panel as the control panel. SLB then replaced the Aussie products in the test panel stores with Aussie products in the new packaging and collected eight weeks of posttest period sales data from stores in both panels.

   BRA's data processing department has produced the file *brughastacked* of the data for this study in a stacked arrangement as shown in the following figure. Marta Davis, Director of BRA's Analytics Division, needs these data to be converted from stacked to unstacked with separate fields for pretest period sales and test period sales for each

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
brughastacked

Stacked Brugha Data

	A	B	C	D
1	Store Number	Panel	Period	Sales
2	96	Control	Pretest	$879.63
3	70	Control	Pretest	$689.72
4	63	Control	Pretest	$638.51
5	19	Control	Pretest	$325.00
6	24	Control	Pretest	$819.76
7	4	Control	Pretest	$942.93
8	81	Control	Pretest	$345.59
9	92	Control	Pretest	$694.70
10	93	Control	Pretest	$299.33
11	50	Control	Pretest	$440.54
12	96	Control	Posttest	$405.26
13	70	Control	Posttest	$270.20
14	86	Control	Posttest	$692.00
15	19	Control	Posttest	$752.90
16	24	Control	Posttest	$662.66
17	4	Control	Posttest	$707.52
18	81	Control	Posttest	$415.85
19	92	Control	Posttest	$894.70
20	93	Control	Posttest	$467.88
21	50	Control	Posttest	$851.67
22	10	Test	Pretest	$635.43
23	36	Test	Pretest	$549.08
24	3	Test	Pretest	$681.12
25	72	Test	Pretest	$389.34
26	52	Test	Pretest	$297.96
27	83	Test	Pretest	$419.57
28	54	Test	Pretest	$619.31
29	94	Test	Pretest	$623.12
30	74	Test	Pretest	$363.88
31	22	Test	Pretest	$980.15
32	10	Test	Posttest	$365.76
33	36	Test	Posttest	$701.63
34	3	Test	Posttest	$993.51
35	72	Test	Posttest	$713.00
36	52	Test	Posttest	$1,093.44
37	83	Test	Posttest	$902.66
38	54	Test	Posttest	$489.67
39	94	Test	Posttest	$803.11
40	74	Test	Posttest	$931.14
41	22	Test	Posttest	$975.83

store and period. Convert these data for Marta Davis so they are unstacked and save the file under the name *broughaunstacked.xlsx*. **LO 3**

5. **Stacking Quarterly Sales Data.** The Accounting Department at van Dyke Industries has produced a data set of the company's quarterly sales revenue for the past five years in an unstacked arrangement. These data are provided in the file *vandykeunstacked* as shown in the following figure.

DATA *file*
vandykeunstacked

	A	B	C	D	E
1		Sales Revenue ($ thousands)			
2	Year	QTR 1	QTR 2	QTR 3	QTR 4
3	1	$562	$577	$578	$526
4	2	$599	$829	$578	$748
5	3	$735	$644	$639	$748
6	4	$785	$870	$861	$526
7	5	$1,228	$1,317	$841	$979

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A financial analyst at van Dyke needs these data stacked so each record corresponds to one year and quarter. Convert these data so they are stacked and save the file under the name *vandykestacked.xlsx*. **LO 3**

6. **Employee Records.** Laghari & Laghari's Personnel Department maintains a file of all employees of the company and each employee's unique Employee ID. The file *llemployees* provides information on employees for Laghari & Laghari's database system, including the first name, last name, and employee ID for each employee in the fields First, Last, and EmpID. The first two digits of the employee ID are the day of the month on which the employee was born, the next two digits are the month in which the employee was born, and the next four digits are the date of the year in which the employee was born. The two digits that immediately follow identify the department in which the employee works, and the last four digits represent other information about the employee. From Noah Chavez's employee ID (14-02-1983-02-4698), we can see that Noah was born on February 14, 1983, and works in department 2.



Laghari & Laghari management wants a new file that includes separate fields for the date of the month on which the employee was born, the month in which the employee was born, the year in which the employee was born, and the department in which the employee works. They also want the First and Last fields combined into a single field called Name with the first and last names separated by a space. Finally, Laghari & Laghari management has determined that the birth years of its current employees range from 1960 to 2003, and it wants to ensure that no employee record has a birth year outside of this range. **LO 4, 5**

   a. Create a new set of data from the file *llemployees* that includes the employee's first name, last name, and each of these four new fields with the field names Day, Month, Year, and Department. Save your file under the name *llemployeeinfo_a.xlsx*.
   b. Using the file *llemployeesinfo_a* that you created in part (a), combine the First and Last fields into a single field FullName with the first and last names separated by a space. Delete the original First and Last fields and save your file under the name *llemployeeinfo_b.xlsx*.
   c. Search the file *llemployeesinfo_b* that you created in part (b) for values for Year that are less than 1960 or greater than 2003 (which Laghari & Laghari has indicated is the range of birth years for its current employees). Provide a list of the names, birthdates, and departments for employees for whom the value of the field Year is outside of this range.

7. **Contest Winners.** Teller Crackers & Snacks (TCS) recently held a contest in which it awarded each of ten randomly drawn entrants a large box of assorted crackers each week for a year. The names and addresses of the ten winners are available in the file *tcswinners*. The fields in the file are First Name, Last Name, Street Address, Apartment number, City, State, and Zip Code.



In order to create mailing labels to use to ship the boxes of crackers to the winners, TCS must create a new file with

   - a field named *Winner* that contains the first and last name separated by a space for each record in the file *tcswinners*
   - a field named Mailing Address that contains the street address, apartment number, city, state, and zip code separated by a comma and a space for each record in the file *tcswinners* (note that the field Apartment Number has legitimate missing values for many records)

   Create these new fields for TCS, delete all other fields, and save the results as a file under the name *tcswinnersmailing.xlsx*. **LO 4**

8. **College Applications.** Bickley Academy is a competitive private college preparatory high school. Bickley's administration is now considering applications from the three local junior high schools: Montgomery, Cleeghorn, and Benson. Bickley's admissions department has compiled a file named *bickley* that contains the first name, last name, email address, and grade point average (GPA) for applicants from these schools.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
bickley

Bickley's Director of Admissions, Joan Bisdubic, would like to look at the relationship between these students' GPAs and the schools they are attending. The school each student is attending can be identified by the domain name in the student's email address. To ensure each applicant receives full and proper consideration, Joan Bisdubic would also like to ensure that the file has the GPA for every applicant. **LO 4, 5**

a. Create a new set of data from the data in the file *bickley* that includes the applicant's first name, last name, email username (the portion of the email address that precedes the @ symbol), email domain (the portion of the email address that follows the @ symbol), and GPA. Save your file under the name *bickleyapp.xlsx*.
b. Using the file *bickleyapp* that you created in part (a), determine if any records have missing values for the GPA field. Provide a list of the names of the applicants for whom the value of the field GPA is missing.

9. **Season Ticket Holders.** The Gulf Shore Blue Crabs minor league baseball team wants to create a new identification system for its season ticket holders using information in the file *bluecrabs*. The file contains the first and last name of the ticket holder and the following information on the ticket holder's seats:

DATA *file*
bluecrabs

- Section
- Level (f for field, l for loge, or b for bleacher)
- Row
- Seat (lowest numbered seat in the block held by the season ticket holder)

The file also includes

- a field named NumSeats that indicates how many seats the season ticket holder has purchased
- a field named Seasons that indicates how many consecutive seasons the season ticket holder has purchased season tickets
- a field named Upgrade that indicates how many seasons have passed since these season tickets have been upgraded
- a field named Request that has a value of Y if the season ticket holder has requested an upgrade for the upcoming season.

The field Upgrade is blank if the season ticket holder's seats have never been upgraded, and the field Request is blank if the season ticket holder has not requested an upgrade for the upcoming season.

Blue Crabs management wants to assign each season ticket holder a unique ID code that comprises the Section, Level, Row, Seat, NumSeats, Seasons, and Upgrade fields with no space between the values. For example, the ID Code for Evander Perez (the first season ticket holder listed in the file *bluecrabs*) should be 3l091496. Blue Crabs management also wants information on all season ticket holders who have requested an upgrade for the upcoming season sorted in descending order by the number of seasons the individual has held season tickets (which is the first criteria for deciding which season ticket holders will receive upgrades). **LO 4, 6**

a. Create a new field named ID Code in the file *bluecrabs* that combines the value of Section, Level, Row, Seat, NumSeats, Seasons, and Upgrade fields with no space between the values for each record as described in the problem. Save your file under the name *bluecrabs_idcodes.xlsx*.
b. Produce a subset of the records in the file *bluecrabs_idcodes* that includes only season ticket holders who have requested upgrades for the upcoming season. Sort these data in descending order by the field Seasons, then provide a list of these season ticket holders and the corresponding values for Season and ID Code in this order.

10. **Conference Registrations.** The North Atlantic Consortium of Nephrology Associations (NACNA) is made up of two professional societies, the North American Nephrologist Society (NANS) and the European Council on Nephrology (ECN). The

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
registrants

NACNA organizes an annual conference, and registration for the conference recently closed. Conference organizers now must print badges for the attendees in preparation for the conference. The NACNA conference registration system produces the file *registrants* that contains the first name, last name, affiliation (NANS or ECN), and membership status (full, student, retired, nonmember) in separate fields for each conference registrant. Note that the Affiliation field should be blank for all registrants with a value of nonmember for membership status.

To print the badges, the first and last names must be combined in a single field called Name and separated by a space. A new field called Color must also be created; the value of this field must be Blue if the record corresponds to a member of the NANS and Red if the record corresponds to a member of the ECN. **LO 4, 5, 7**

a. Using the data in the file *registrants*, create a new called Name that contains the first and last names separated by a space for each registrant.
b. In the data you created in part (a), add a new field called Color that has the value Blue if the record corresponds to a member of the NANS and Red if the record corresponds to a member of the ECN. Save your file under the name *badges.xlsx*.
c. Using the file *badges* that you created in part (b), determine if any records are missing values for the Association field. Provide a list of the names of the registrants for whom membership status is full, student, or retired and the value of the field Association is missing.

11. **State and Regional Advertising Sales.** *Valley of the Grape* is a show about wine and winemaking that is available on streaming services throughout the United States. Kleskin & Relouk (K&R), producers of *Valley of the Grape*, generate revenue from sales of advertising on the show. Advertising sales by state and region (Northeast, South, Midwest, West) for the previous year are available in the file *krsales*. K&R management wants to receive reports of state-by-state sales for each region. **LO 6**

DATA*file*
krsales

a. Produce a subset of the records in the file *krsales* that include only states in the Northeast region. Save your results as a file under the name *krsales_ne.xlsx*.
b. Produce a subset of the records in the file *krsales* that include only states in the South region. Save your results as a file under the name *krsales_s.xlsx*.
c. Produce a subset of the records in the file *krsales* that include only states in the Midwest region. Save your results as a file under the name *krsales_mw.xlsx*.
d. Produce a subset of the records in the file *krsales* that include only states in the West region. Save your results as a file under the name *krsales_w.xlsx*.

12. **Textbook Sales.** Bustamante & Sons (B&S), Inc. publishes a small line of textbooks for introductory undergraduate business courses. Deron Ackerman, Director of Sales for B&S, is now performing annual reviews of the members of the B&S sales staff. Ackerman is reviewing the sales made by Beverly Harrelson for the past year, which are provided in the file *harrelsonsales*. Ackerman would like to review Harrelson's sales performance, specifically for Bartram University and for the book with product code 7314. **LO 6**

DATA*file*
harrelsonsales

a. Produce a subset of the records in the file *harrelsonsales* that includes only records for Bartram University. Save your results as a file under the name *harrelsonbartram.xlsx*.
b. Produce a subset of the records in the file *harrelsonsales* that includes only records for the book with product code 7314. Save your results as a file under the name *harrelson7314sales.xlsx*.

13. **Lighting Fixture Prices and Sales.** Tomlinson's Lumens, Inc. (TLI) is a large manufacturer of lighting fixtures and parts. TLA's Accounting department receives files *tomlinsonprices* and *tomlinsonunits* that summarize the price and units sold, respectively, for each product it produces on a quarterly basis.

DATA*file*
tomlinsonprices, tomlinsonunits

Use this information to calculate the company's most recent quarterly total sales for TLI's Accounting department. **LO 7, 8**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



khatripopulation, khatrisales

14. **Dollar Sales and Population by Market.** Donald Midkif, a data scientist with Khatri, Inc. has been tasked with determining in which of the largest U.S. cities the products Khatri manufactures and distributes are most popular as measured by per capita sales. Midkif has access to two files, *khatripopulation* and *khatrisales* that provide the population and Khatri dollar sales, respectively, for each of the fifty largest U.S. cities for last year.

    Use this information to calculate Khatri's per capita sales (dollar sales divided by population) in each of these fifty cities for last year for Midkif. **LO 7, 8**



primoappts, primolenses

15. **Eyewear Customer Data.** A local PrimoVision Eyewear store maintains a file named *primoappts* that contains the customer ID number and date of most recent appointment for each of its customers. The store also maintains a file named *primolenses* that contains the customer ID number and type of corrective lenses each customer currently uses. The types of corrective lenses includes:

    - monofocal lenses
    - bifocal lenses
    - trifocal lenses
    - progressive lenses
    - disposable contact lenses
    - daily wear contact lenses
    - extended wear contact lenses
    - bifocal contact lenses
    - rigid gas permeable contact lenses
    - toric contact lenses

    The office manager of the PrimoVision Eyewear store needs a file that contains the fields customer ID number, date of most recent visit for each of its customers, and type of corrective lenses the customer currently uses. The manager is also concerned that information on the most recent visit or type of corrective lenses the customer currently uses are missing for some records, and they suspect that these data also include some duplicate records. **LO 5, 8**

    a. Append the appropriate value for the field Lens Type in the file *primolenses* to each record in the file *primoappts*. Save your results as a file under the name *primodata.xlsx*.
    b. Search the file *primodata* that you created in part (a) for missing values for the Lens Type field. Provide a list of the names of the customers for whom the value of the field Lens Type is missing.
    c. Search the file *primodata* that you created in part (a) for duplicate records. Provide a list of the customer names for the duplicate records you find.



sba494destination, sba494class

16. **Passenger Records.** Sweet Bayou Airlines (SBA) flight 494 departs daily from Shreveport for Nashville at 8:47 a.m. The manager of the SBA office in Shreveport must receive a file that combines to contents of the file *sba494destination*, which contains the confirmation number and final destination for each passenger, and the file *sba494class*, which contains the confirmation number and class of seat (first, business, or coach) for each passenger. These combined data must also be searched for duplicate records to ensure no passenger is booked in multiple seats for the flight. **LO 5, 8**

    a. Append the appropriate value for the field Destination in the file *sba494destination* to each record in the file *sba494class*. Save your results as a file under the name *sba494final.xlsx*.
    b. Search the file *sba494final* that you created in part (a) for duplicate records. Provide a list of the passenger names, destinations, seat, class, and confirmation for the duplicate records you find.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

17. **Flight Passenger Class.** A data scientist at Sweet Bayou Airlines (SBA) wants to analyze data related to SBA's flight 494, which departs daily from Shreveport for Nashville at 8:47 a.m. To facilitate this analysis, the data scientist needs the data to include a dummy variable that is equal to 1 if the passenger has a seat in First Class and 0 otherwise, a dummy variable that is equal to 1 if the passenger has a seat in Business Class and 0 otherwise, and a dummy variable that is equal to 1 if the passenger has a seat in Coach Class and 0 otherwise. Using the file *sba494class*, create the three dummy variables in new fields and name these fields FCDummy, BCDummy, and CCDummy. Save your results as a file under the name *sba494classdummy.xlsx*. **LO 7**



sba494class

18. **Iowa Housing Prices**. A realtor who flips houses (purchases and remodels them with the goal of reselling them for profit) has collected data on a variety of house characteristics and selling prices for a sample of 1,198 recent house sales in Iowa (Source: *Journal of Statistics Education*). For each house sold, the following data are provided in the file *iowahouses*.

    - LotArea = Lot size in square feet
    - CentralAir = 1 has central air-conditioning, 0 does not have central air-conditioning
    - GrLivArea = Above grade (ground) living area in square feet
    - FullBath = Full bathrooms above ground
    - HalfBath = Half bathrooms above ground
    - BedroomAbvGr = Bedrooms above ground
    - GarageCars = Size of the garage in car capacity
    - Age = Age of the house when sold (in years)
    - SalePrice = Selling price of the house



iowahouses

The realtor would like to add a field called PricePerSQFT that contains the sales price per square foot (the value of the field SalePrice divided by the value of the field GrLivArea for each record) and a field called OldOrNew that has a value of Old if the age of the home is at least 20 years and has a value of New otherwise for each record. **LO 7**

a. Create a new field named PricePerSQFT in the file *iowahouses*. Define the value of this field for each record to equal the field's SalePrice divided by the field GrLivArea.
b. Add a new field named OldOrNew to your results in part (a). Define the value of this field for each record to equal Old if the age of the home is at least 20 years, and New otherwise. Save your results as a file under the name *iowahousesrevised.xlsx*.

19. **Duration of Unemployment**. A study on factors that may be related to the number of weeks a person has been jobless has been conducted in Cheyenne and Memphis. The data collected in Cheyenne are provided in the file *cheyenneunemp*, and the data collected in Memphis are provided in the file *memphisunemp*. The following information was collected in the study.

    - Weeks — The number of weeks a person has been jobless due to a layoff.
    - Age — The age of the person
    - Educ — The number of years of education
    - Married — 1 if married, 0 otherwise
    - Primary — 1 if the primary income earner, 0 otherwise
    - Tenure — The number of years on the previous job
    - Manager — 1 if previous job was classified as a management occupation, 0 otherwise
    - Sales — 1 if previous job was classified as a sales occupation, 0 otherwise
    - City — City in which the data were collected



cheyenneunemp, memphisunemp

The data from these two files must be combined into a single file to facilitate their analysis. Append the 50 records in the file *cheyenneunemp* to the 68 records in the file

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

memphisunemp. Save your results as a file under the name *combinedunepm.xlsx*. **LO 8**

20. **Risk of Stroke**. In a recent study on risk factors for strokes, the American Heart Association (AHA) collected data on age, systolic blood pressure, diastolic blood pressure, and risk of stroke; risk is interpreted as the probability (times 100) that a person will have a stroke over the next 10-year period. Data collected from smokers is provided in the file *ahasmokers*, and data collected from nonsmokers is provided in the file *ahanonsmokers*.

DATA*file*
ahasmokers, ahanonsmokers

These data must be combined in order to facilitate their analysis. Systolic blood pressure outside of the range of 80–200 and diastolic blood pressure outside of the range of 60–130 are extremely rare, so the data must also be reviewed to ensure there are no records with a systolic blood pressure outside of the range 80–200 and no diastolic blood pressure is outside of the range of 60–140. Finally, the AHA wants to alert all patients who have severe hypertension, which is characterized by a systolic blood pressure above 180, so the organization wants to add a field to the final data that indicates whether a subject has severe hypertension. **LO 5, 7**

a. Create a new field named Smoker? in the file *ahasmokers*. Define the value of this field for each record as 1.
b. Create a new field named Smoker? in the file *ahanonsmokers*. Define the value of this field for each record as 0.
c. Combine all records from these two files into a single file. Save the file under the name *ahastudy.xlsx*.
d. Add a new field named SevereHyp to your results in part (c). Define the value of this field for each record to equal Yes if the systolic blood pressure is above 180 and No otherwise. Save your results as a file under the name *ahaalert.xlsx*.
e. Identify all records for which the systolic blood pressure is outside the range 80–210. Provide a list of the subject numbers for all records for which the systolic blood pressure is outside of this range.
f. Identify all records for which the diastolic blood pressure is outside of the range of 60–140. Provide a list of the subject numbers for all records for which the diastolic blood pressure is outside of this range.

Case Problem 1: Usman Solutions

Usman Solutions is a national marketing research firm that specializes in consumer survey research. Usman was recently awarded a large study of consumer attitudes on the new salted caramel flavor of Crisp Spuds brand potato chips that has been developed by Porter & Hinkle, the manufacturer and distributor of Crisp Spuds. Porter & Hinkle hopes that the new salted caramel flavor of Crisp Spuds will appeal to young consumers who do not regularly consume Crisp Spuds products.

The questionnaire designed by Usman was administered to respondents in Knoxville, Tennessee and Boise, Idaho, the two U.S. cities Porter & Hinkle believes are most representative of the market for Crisp Spuds. In addition to asking for the age of the respondent and whether the respondent regularly consumes any Crispy Spuds product, the survey asks the respondent to taste a sample and rate the crispiness, saltiness, sweetness, flavor, texture, and value of the new salted caramel flavor of Crisp Spuds. Respondents were also asked to provide their overall rating of the new salted caramel flavor of Crisp Spuds. Each rating question utilized a ten-point scale in which 1 was the lowest possible rating and 10 was the highest possible rating.

DATA*file*
knoxville, boise

Usman has received two tab delimited files, *knoxville.txt* and *boise.txt*, that contain the survey responses collected in Knoxville and Boise, respectively. Ana Mendoza, Usman's Manager of Data Integrity, is now responsible for supervising the wrangling of these raw data in preparation for their analysis by Usman's Analytics Division. The field names in these two data sets are Respondent, Age, Regularly Consume Crispy Spuds?, Crispiness, Saltiness, Sweetness, Flavor, Texture, Value, and Overall. **LO 2, 5, 7**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Managerial Report

1. Import the data in the files *knoxville.txt* and *boise.txt* into Excel, add a field named Market to each set of data, and fill this new field in each set of data with the name of the respective market (Knoxville and Boise). Combine these resulting data into a single set of data and save the data with the name *porter_hinkle.xlsx*.
2. Count the number of missing values for fields Respondent, Age, Regularly Consume Crispy Spuds?, Crispiness, Saltiness, Sweetness, Flavor, Texture, Value, and Overall in the set of data created in part (1). Create a report that shows the market, respondent number, and field(s) with missing values for each record you identify.
3. Count and identify all duplicate records in the set of data created in part (1). Create a list of the market and respondent number for each set of duplicate records you identify so these records can be investigated and considered for deletion.
4. Porter & Hinkle's Analytics Division wants to use a new field called Taste that is the mean of the fields Saltiness, Sweetness, and Flavor. Generate this new field for the set of data you created in part (1).
5. In response to Porter & Hinkle's hope that the new salted caramel flavor of Crisp Spuds will appeal to young consumers who do not regularly consume Crisp Spuds products, create a new set of data that contains all respondents under 35 years old who responded No to the question on whether they regularly consume Crispy Spuds from the set of data you created in part (1) and enhanced in part (4). Save this set of data under the name *young_no.xlsx*.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 5

Probability: An Introduction to Modeling Uncertainty

Contents

Analytics in Action: National Aeronautics and Space Administration

5.1 Events and Probabilities

5.2 Some Basic Relationships of Probability
- Complement of an Event
- Addition Law

5.3 Conditional Probability
- Independent Events
- Multiplication Law
- Bayes' Theorem

5.4 Random Variables
- Discrete Random Variables
- Continuous Random Variables

5.5 Discrete Probability Distributions
- Custom Discrete Probability Distribution
- Expected Value and Variance
- Discrete Uniform Probability Distribution
- Binomial Probability Distribution
- Poisson Probability Distribution

5.6 Continuous Probability Distributions
- Uniform Probability Distribution
- Triangular Probability Distribution
- Normal Probability Distribution
- Exponential Probability Distribution

Summary 240

Glossary 240

Problems 242

Available in the Cengage eBook:
- Appendix: Discrete Probability Distributions with R
- Appendix: Continuous Probability Distributions with R

Learning Objectives

After completing this chapter, you will be able to

LO 1 Calculate the number of outcomes for random experiments and assign probabilities to outcomes and events.

LO 2 Calculate and interpret the probability of the complement, the union, and the intersection of events.

LO 3 Calculate and interpret the conditional probability associated with two events.

LO 4 Identify and interpret mutually exclusive events and independent events.

LO 5 Create and interpret joint probability tables.

LO 6 Identify and interpret prior probabilities and posterior probabilities; apply Bayes' theorem to calculate posterior probabilities.

LO 7 Develop valid probability distributions for discrete random variables, and use these distributions to calculate and interpret the expected value, variance and standard deviation.

LO 8 Calculate and interpret the expected value, variance, and standard deviation for discrete random variables.

LO 9 Identify and use the binomial probability function to compute the probability of a specified number of successes out of a given number of trials and the expected value, variance, and standard deviation.

LO 10 Identify and use the Poisson probability function to compute the probability of a specified number of occurrences over a given interval and calculate the expected value, variance, and standard deviation of the number of occurrences.

LO 11 Identify and use the uniform probability distribution to compute the probability that a random variable assumes a value within a specified interval, and calculate the expected value, variance, and standard deviation for the random variable.

LO 12 Identify and use the triangular probability distribution to compute the probability that a random variable assumes a value within a specified interval.

LO 13 Identify and use the normal probability distribution to compute the probability that a random variable assumes a value within a specified interval.

LO 14 Calculate and interpret the value of a random variable corresponding to a specified probability for a normal probability distribution.

LO 15 Identify and use the exponential probability distribution to compute the probability that a random variable assumes a value within a specified interval.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

National Aeronautics and Space Administration*

Washington, D.C.

The National Aeronautics and Space Administration (NASA) is the U.S. government agency that is responsible for the U.S. civilian space program and for aeronautics and aerospace research. NASA is best known for its manned space exploration; its mission statement is to "drive advances in science, technology, aeronautics, and space exploration to enhance knowledge, education, innovation, economic vitality and stewardship of Earth." With more than 17,000 employees, NASA oversees many different space-based missions including work on the International Space Station, exploration beyond our solar system with the Hubble and James Webb telescopes, and planning for possible future astronaut missions to the moon and Mars.

Although NASA's primary mission is space exploration, its expertise has been called on in assisting countries and organizations throughout the world in nonspace endeavors. In one such situation, the San José copper and gold mine in Copiapó, Chile, caved in, trapping 33 men more than 2,000 feet underground. It was important to bring the men safely to the surface as quickly as possible, but it was also imperative that the rescue effort be carefully designed and implemented to save as many miners as possible. The Chilean government asked NASA to provide assistance in developing a rescue method. NASA sent a four-person team consisting of an engineer with expertise in vehicle design, two physicians, and a psychologist with knowledge about issues of long-term confinement.

The probability of success and the failure of various other rescue methods was prominent in the thoughts of everyone involved. Since no historical data were available to apply to this unique rescue situation, NASA scientists developed subjective probability estimates for the success and failure of various rescue methods based on similar circumstances experienced by astronauts returning from short- and long-term space missions. The probability estimates provided by NASA guided officials in the selection of a rescue method and provided insight as to how the miners would survive the ascent in a rescue cage. The rescue method designed by the Chilean officials in consultation with the NASA team resulted in the construction of 13-foot-long, 924-pound steel rescue capsule that would be used to bring up the miners one at a time. All miners were rescued, with the last emerging 68 days after the cave-in occurred.

In this chapter, you will learn about probability as well as how to compute and interpret probabilities for a variety of situations. The basic relationships of probability, conditional probability, and Bayes' theorem will be covered. We will also discuss the concepts of random variables and probability distributions and illustrate the use of some of the more common discrete and continuous probability distributions.

*The authors are indebted to Dr. Michael Duncan and Clinton Cragg at NASA for providing this Analytics in Action.

Uncertainty is an ever-present fact of life for decision makers, and much time and effort are spent trying to plan for, and respond to, uncertainty. Consider the CEO who has to make decisions about marketing budgets and production amounts using forecasted demands. Or consider the financial analyst who must determine how to build a client's portfolio of stocks and bonds when the rates of return for these investments are not known with certainty. In many business scenarios, data are available to provide information on possible outcomes for some decisions, but the exact outcome from a given decision is almost never known with certainty because many factors are outside the control of the decision maker (e.g., actions taken by competitors, the weather).

Probability is the numerical measure of the likelihood that an event will occur.[1] Therefore, it can be used as a measure of the uncertainty associated with an event.

[1]Note that there are several different possible definitions of probability, depending on the method used to assign probabilities. This includes the classical definition, the relative frequency definition, and the subjective definition of probability. In this text, we most often use the relative frequency definition of probability, which assumes that probabilities are based on empirical data. For a more thorough discussion of the different possible definitions of probability see Chapter 4 of Camm, et al., *An Introduction to Statistics for Business & Economics*, 15e (2023).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This measure of uncertainty is often communicated through a probability distribution. Probability distributions are extremely helpful in providing additional information about an event, and as we will see in later chapters in this textbook, they can be used to help a decision maker evaluate possible actions and determine the best course of action.

5.1 Events and Probabilities

In discussing probabilities, we start by defining a **random experiment** as a process that generates well-defined outcomes. Several examples of random experiments and their associated outcomes are shown in Table 5.1.

By specifying all possible outcomes, we identify the **sample space** for a random experiment. Consider the first random experiment in Table 5.1—a coin toss. The possible outcomes are head and tail. If we let S denote the sample space, we can use the following notation to describe the sample space.

$$S = \{\text{Head, Tail}\}$$

Suppose we consider the second random experiment in Table 5.1—rolling a die. The possible experimental outcomes, defined as the number of dots appearing on the upward face of the die, are the six points in the sample space for this random experiment.

$$S = \{1, 2, 3, 4, 5, 6\}$$

Outcomes and events form the foundation of the study of probability. Formally, an **event** is defined as a collection of outcomes. For example, consider the case of an expansion project being undertaken by California Power & Light Company (CP&L). CP&L is starting a project designed to increase the generating capacity of one of its plants in Southern California. An analysis of similar construction projects indicates that the possible completion times for the project are 8, 9, 10, 11, and 12 months. Each of these possible completion times represents a possible outcome for this project. Table 5.2 shows the number of past construction projects that required 8, 9, 10, 11, and 12 months.

Let us assume that the CP&L project manager is interested in completing the project in 10 months or less. Referring to Table 5.2, we see that three possible outcomes (8 months, 9 months, and 10 months) provide completion times of 10 months or less. Letting C denote the event that the project is completed in 10 months or less, we write:

$$C = \{8, 9, 10\}$$

Event C is said to occur if *any one* of these outcomes occurs.

Table 1.1 Random Experiments and Experimental Outcomes

Random Experiment	Experimental Outcomes
Toss a coin	Head, tail
Roll a die	1, 2, 3, 4, 5, 6
Conduct a sales call	Purchase, no purchase
Hold a particular share of stock for one year	Price of stock goes up, price of stock goes down, no change in stock price
Reduce price of product	Demand goes up, demand goes down, no change in demand

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 5.2 Completion Times for 40 CP&L Projects

Completion Time (months)	No. of Past Projects Having This Completion Time	Probability of Outcome
8	6	6/40 = 0.15
9	10	10/40 = 0.25
10	12	12/40 = 0.30
11	6	6/40 = 0.15
12	6	6/40 = 0.15
Total	40	1.00

A variety of additional events can be defined for the CP&L project:

L = The event that the project is completed in *less than* 10 months = $\{8, 9\}$

M = The event that the project is completed in *more than* 10 months = $\{11, 12\}$

In each case, the event must be identified as a collection of outcomes for the random experiment.

The **probability of an event** is equal to the sum of the probabilities of outcomes for the event. Using this definition and given the probabilities of outcomes shown in Table 5.2, we can now calculate the probability of the event $C = \{8, 9, 10\}$. The probability of event C, denoted $P(C)$, is given by

$$P(C) = P(8) + P(9) + P(10) = 0.15 + 0.25 + 0.30 = 0.70$$

Similarly, because the event that the project is completed in less than 10 months is given by $L = \{8, 9\}$, the probability of this event is given by

$$P(L) = P(8) + P(9) = 0.15 + 0.25 = 0.40$$

Finally, for the event that the project is completed in more than 10 months, we have $M = \{11, 12\}$ and thus

$$P(M) = P(11) + P(12) = 0.15 + 0.15 = 0.30$$

Using these probability results, we can now tell CP&L management that there is a 0.70 probability that the project will be completed in 10 months or less, a 0.40 probability that it will be completed in less than 10 months, and a 0.30 probability that it will be completed in more than 10 months.

5.2 Some Basic Relationships of Probability

Complement of an Event

The complement of event A is sometimes written as $\bar{A}$ or A′ in other textbooks.

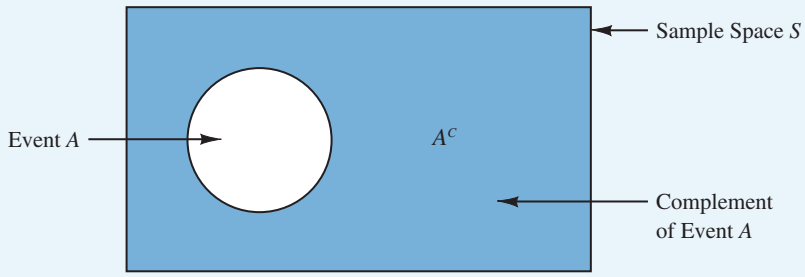
Given an event A, the **complement of A** is defined to be the event consisting of all outcomes that are *not* in A. The complement of A is denoted by A^C. Figure 5.1 shows what is known as a **Venn diagram**, which illustrates the concept of a complement. The rectangular area represents the sample space for the random experiment and, as such, contains all possible outcomes. The circle represents event A and contains only the outcomes that belong to A. The shaded region of the rectangle contains all outcomes not in event A and is by definition the complement of A.

In any probability application, either event A or its complement A^C must occur. Therefore, we have

$$P(A) + P(A^C) = 1$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.1 Venn Diagram for Event A



Solving for $P(A)$, we obtain the following result:

Computing Probability Using the Complement

$$P(A) = 1 - P(A^C) \tag{5.1}$$

Equation (5.1) shows that the probability of an event A can be computed easily if the probability of its complement, $P(A^C)$, is known.

As an example, consider the case of a sales manager who, after reviewing sales reports, states that 80% of new customer contacts result in no sale. By allowing A to denote the event of a sale and A^C to denote the event of no sale, the manager is stating that $P(A^C) = 0.80$. Using equation (5.1), we see that

$$P(A) = 1 - P(A^C) = 1 - 0.80 = 0.20$$

We can conclude that a new customer contact has a 0.20 probability of resulting in a sale.

Addition Law

The addition law is helpful when we are interested in knowing the probability that at least one of two events will occur. That is, with events A and B we are interested in knowing the probability that event A or event B occurs or both events occur.

Before we present the addition law, we need to discuss two concepts related to the combination of events: the *union* of events and the *intersection* of events. Given two events A and B, the **union of A and B** is defined as the event containing all outcomes belonging to A or B or both. The union of A and B is denoted by $A \cup B$.

The Venn diagram in Figure 5.2 depicts the union of A and B. Note that one circle contains all the outcomes in A and the other all the outcomes in B. The fact that the circles overlap indicates that some outcomes are contained in both A and B.

The definition of the **intersection of A and B** is the event containing the outcomes that belong to both A and B. The intersection of A and B is denoted by $A \cap B$. The Venn diagram depicting the intersection of A and B is shown in Figure 5.3. The area in which the two circles overlap is the intersection; it contains outcomes that are in both A and B.

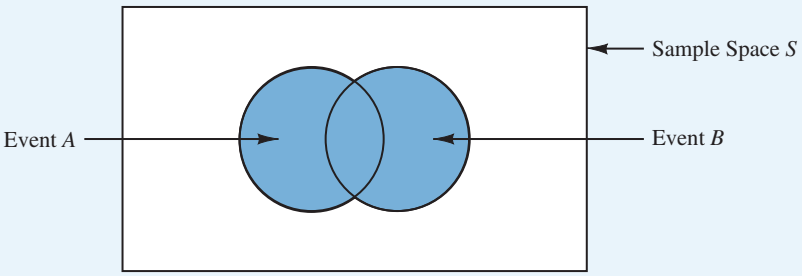
The **addition law** provides a way to compute the probability that event A or event B occurs or both events occur. In other words, the addition law is used to compute the probability of the union of two events. The addition law is written as follows:

Addition Law

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{5.2}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.2 Venn Diagram for the Union of Events *A* and *B*



To understand the addition law intuitively, note that the first two terms in the addition law, $P(A) + P(B)$, account for all the sample points in $A \cup B$. However, because the sample points in the intersection $A \cap B$ are in both A and B, when we compute $P(A) + P(B)$, we are in effect counting each of the sample points in $A \cap B$ twice. We correct for this double counting by subtracting $P(A \cap B)$.

As an example of the addition law, consider a study conducted by the human resources manager of a major computer software company. The study showed that 30% of the employees who left the firm within two years listed dissatisfaction with their salary as one of the reasons, 20% listed dissatisfaction with their work assignments as one of the reasons, and 12% of the former employees indicated dissatisfaction with *both* their salary and their work assignments. What is the probability that an employee who leaves within two years does so because of dissatisfaction with salary, dissatisfaction with the work assignment, or both?

Let

S = the event that the employee leaves because of salary
W = the event that the employee leaves because of work assignment

We can also think of this probability in the following manner: What proportion of employees either left because of salary or left because of work assignment?

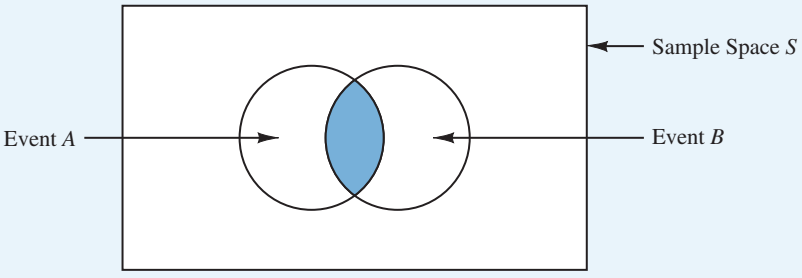
From the survey results, we have $P(S) = 0.30$, $P(W) = 0.20$, and $P(S \cap W) = 0.12$. Using the addition law from equation (5.2), we have

$$P(S \cup W) = P(S) + P(W) - P(S \cap W) = 0.30 + 0.20 - 0.12 = 0.38$$

This calculation tells us that there is a 0.38 probability that an employee will leave for salary or work assignment reasons.

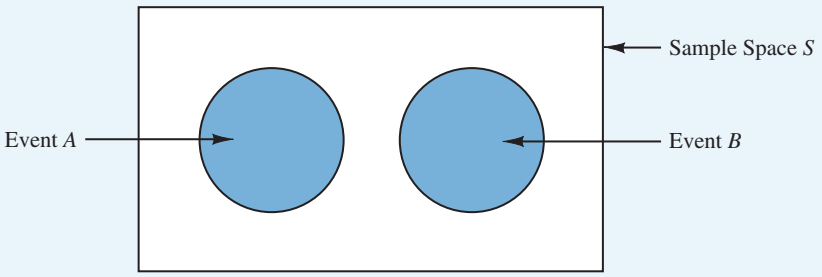
Before we conclude our discussion of the addition law, let us consider a special case that arises for **mutually exclusive events**. Events A and B are mutually exclusive if the

Figure 5.3 Venn Diagram for the Intersection of Events *A* and *B*



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.4 Venn Diagram for Mutually Exclusive Events



occurrence of one event precludes the occurrence of the other. Thus, a requirement for A and B to be mutually exclusive is that their intersection must contain no sample points. The Venn diagram depicting two mutually exclusive events A and B is shown in Figure 5.4. In this case $P(A \cap B) = 0$ and the addition law can be written as follows:

Addition Law for Mutually Exclusive Events

$$P(A \cup B) = P(A) + P(B)$$

More generally, two events are said to be mutually exclusive if the events have no outcomes in common.

Notes + Comments

1. The concept of identifying uncertainty in data was introduced in Chapters 2 and 3 through descriptive statistics and data visualization techniques, respectively. This chapter expands on our discussion of modeling uncertainty by formalizing the concept of probability and introducing the concept of probability distributions.
2. The addition law can be extended beyond two events. For example, the addition law for three events A, B, and C is $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$. Similar logic can be used to derive the expressions for the addition law for more than three events.

5.3 Conditional Probability

Often, the probability of one event is dependent on whether some related event has already occurred. Suppose we have an event A with probability $P(A)$. If we learn that a related event, denoted by B, has already occurred, we take advantage of this information by calculating a new probability for event A. This new probability of event A is called a **conditional probability** and is written $P(A \mid B)$. The notation $\mid$ indicates that we are considering the probability of event A *given* the condition that event B has occurred. Hence, the notation $P(A \mid B)$ reads "the probability of A given B."

To illustrate the idea of conditional probability, consider a bank that is interested in the mortgage default risk for its home mortgage customers. Table 5.3 shows the first 25 records of the 300 home mortgage customers at Lancaster Savings and Loan, a company that specializes in high-risk subprime lending. Some of these home mortgage customers have defaulted on their mortgages and others have continued to make on-time payments. These data include the age of the customer at the time of mortgage origination, the marital status of the customer (single or married), the annual income of the customer, the mortgage amount, the number of payments made by the customer per year on the mortgage, the total amount paid by the customer over the lifetime of the mortgage, and whether or not the customer defaulted on their mortgage.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 5.3 Subset of Data from 300 Home Mortgages of Customers at Lancaster Savings and Loan

Customer No.	Age	Marital Status	Annual Income ($)	Mortgage Amount ($)	Payments per Year	Total Amount Paid ($)	Default on Mortgage?
1	37	Single	172,125.70	473,402.96	24	581,885.13	Yes
2	31	Single	108,571.04	300,468.60	12	489,320.38	No
3	37	Married	124,136.41	330,664.24	24	493,541.93	Yes
4	24	Married	79,614.04	230,222.94	24	449,682.09	Yes
5	27	Single	68,087.33	282,203.53	12	520,581.82	No
6	30	Married	59,959.80	251,242.70	24	356,711.58	Yes
7	41	Single	99,394.05	282,737.29	12	524,053.46	No
8	29	Single	38,527.35	238,125.19	12	468,595.99	No
9	31	Married	112,078.62	297,133.24	24	399,617.40	Yes
10	36	Single	224,899.71	622,578.74	12	1,233,002.14	No
11	31	Married	27,945.36	215,440.31	24	285,900.10	Yes
12	40	Single	48,929.74	252,885.10	12	336,574.63	No
13	39	Married	82,810.92	183,045.16	12	262,537.23	No
14	31	Single	68,216.88	165,309.34	12	253,633.17	No
15	40	Single	59,141.13	220,176.18	12	424,749.80	No
16	45	Married	72,568.89	233,146.91	12	356,363.93	No
17	32	Married	101,140.43	245,360.02	24	388,429.41	Yes
18	37	Married	124,876.53	320,401.04	4	360,783.45	Yes
19	32	Married	133,093.15	494,395.63	12	861,874.67	No
20	32	Single	85,268.67	159,010.33	12	308,656.11	No
21	37	Single	92,314.96	249,547.14	24	342,339.27	Yes
22	29	Married	120,876.13	308,618.37	12	472,668.98	No
23	24	Single	86,294.13	258,321.78	24	380,347.56	Yes
24	32	Married	216,748.68	634,609.61	24	915,640.13	Yes
25	44	Single	46,389.75	194,770.91	12	385,288.86	No

Lancaster Savings and Loan is interested in whether the probability of a customer defaulting on a mortgage differs by marital status. Let

S = event that a customer is single
M = event that a customer is married
D = event that a customer defaulted on their mortgage
D^C = event that a customer did not default on their mortgage

Table 5.4 shows a crosstabulation for two events that can be derived from the Lancaster Savings and Loan mortgage data.

Chapter 3 discusses PivotTables in more detail.

Note that we can easily create Table 5.4 in Excel using a PivotTable by using the following steps:

DATA*file*
mortgagedefault

Step 1. Click the **Insert** tab on the Ribbon
Step 2. Click **PivotTable** in the **Tables** group
Step 3. When the **Create PivotTable** dialog box appears:
Choose **Select a Table or Range**
Enter *A1:H301* in the **Table/Range:** box

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Select **New Worksheet** as the location for the PivotTable Report
Click **OK**

Step 4. In the **PivotTable Fields** area go to **Drag fields between areas below:**
Drag the **Marital Status** field to the **Rows** area
Drag the **Default on Mortgage?** field to the **Columns** area
Drag the **Customer Number** field to the **Values** area

Step 4. Click on **Sum of Customer Number** in the **Values** area and select **Value Field Settings**

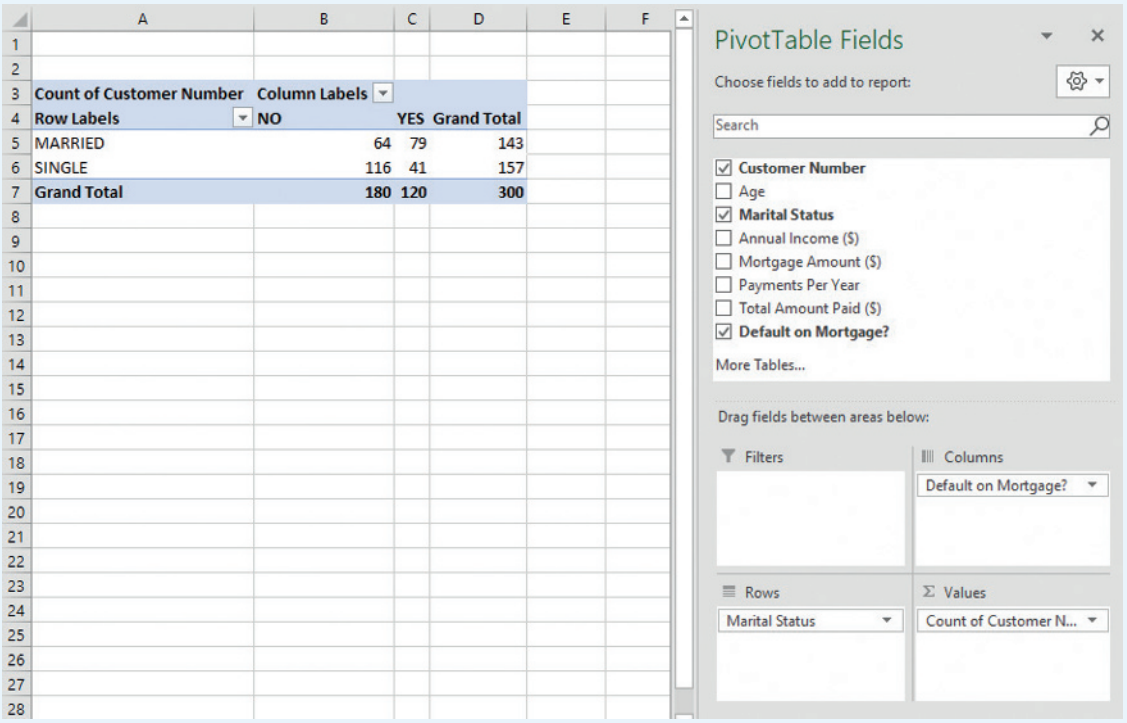
Step 6. When the **Value Field Settings** dialog box appears:
Under **Summarize value field by**, select **Count** and click **OK**

These steps produce the PivotTable shown in Figure 5.5.

Table 5.4 Crosstabulation of Marital Status and if Customer Defaults on Mortgage

Marital Status	No Default	Default	Total
Married	64	79	143
Single	116	41	157
Total	*180*	*120*	*300*

Figure 5.5 PivotTable for Marital Status and Whether Customer Defaults on Mortgage



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

From Table 5.4 or Figure 5.5, the probability that a customer defaults on their mortgage is $120/300 = 0.4$. The probability that a customer does not default on their mortgage is $1 - 0.4 = 0.6$ (or $180/300 = 0.6$). But is this probability different for married customers as compared with single customers? Conditional probability allows us to answer this question.

We can also think of this joint probability in the following manner: What proportion of all customers is both married and defaulted on their loans?

But first, let us answer a related question: What is the probability that a randomly selected customer does not default on their mortgage and the customer is married? The probability that a randomly selected customer is married and the customer defaults on their mortgage is written as $P(M \cap D)$. This probability is calculated as $P(M \cap D) = \frac{79}{300} = 0.2633$.

Similarly,

$P(M \cap D^C) = \frac{64}{300} = 0.2133$ is the probability that a randomly selected customer is married and that the customer does not default on their mortgage.
$P(S \cap D) = \frac{41}{300} = 0.1367$ is the probability that a randomly selected customer is single and that the customer defaults on their mortgage.
$P(S \cap D^C) = \frac{116}{300} = 0.3867$ is the probability that a randomly selected customer is single and that the customer does not default on their mortgage.

Because each of these values gives the probability of the intersection of two events, these probabilities are called **joint probabilities**. Table 5.5, which provides a summary of the probability information for customer defaults on mortgages, is referred to as a joint probability table.

The values in the Total column and Total row (the margins) of Table 5.5 provide the probabilities of each event separately. That is, $P(M) = 0.4766$, $P(S) = 0.5234$, $P(D^C) = 0.6000$, and $P(D) = 0.4000$. These probabilities are referred to as **marginal probabilities** because of their location in the margins of the joint probability table. The marginal probabilities are found by summing the joint probabilities in the corresponding row or column of the joint probability table. From the marginal probabilities, we see that 60% of customers do not default on their mortgage, 40% of customers default on their mortgage, 47.66% of customers are married, and 52.34% of customers are single.

Let us begin the conditional probability analysis by computing the probability that a customer defaults on their mortgage given that the customer is married. In conditional probability notation, we are attempting to determine $P(D \mid M)$, which is read as "the probability that the customer defaults on the mortgage given that the customer is married." To calculate $P(D \mid M)$, first we note that we are concerned only with the 143 customers who are married (M). Because 79 of the 143 married customers defaulted on their mortgages, the probability of a customer defaulting given that the customer is married is $79/143 = 0.5524$.
In other words, given that a customer is married, there is a 55.24% chance that they will default. Note also that the conditional probability $P(D \mid M)$ can be computed as the ratio of the joint probability $P(D \cap M)$ to the marginal probability $P(M)$.

$$P(D \mid M) = \frac{P(D \cap M)}{P(M)} = \frac{0.2633}{0.4766} = 0.5524$$

We can use the PivotTable from Figure 5.5 to easily create the joint probability table in Excel. To do so, right-click on any of the numerical values in the PivotTable, select ***Show Values As****, and choose* ***% of Grand Total****. The resulting values, which are percentages of the total, can then be divided by 100 to create the probabilities in the joint probability table.*

Table 5.5 Joint Probability Table for Customer Mortgage Defaults

Joint Probabilities

	No Default (D^C)	Default (D)	Total
Married (M)	0.2133	0.2633	0.4766
Single (S)	0.3867	0.1367	0.5234
Total	*0.6000*	*0.4000*	*1.0000*

Marginal Probabilities

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The fact that a conditional probability can be computed as the ratio of a joint probability to a marginal probability provides the following general formula for conditional probability calculations for two events A and B.

Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{5.3}$$

or

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{5.4}$$

We have already determined that the probability a customer who is married will default is 0.5524. How does this compare to a customer who is single? That is, we want to find $P(D \mid S)$. From equation (5.3), we can compute $P(D \mid S)$ as

$$P(D \mid S) = \frac{P(D \cap S)}{P(S)} = \frac{0.1367}{0.5234} = 0.2611$$

In other words, the chance that a customer will default if the customer is single is 26.11%. This is substantially less than the chance of default if the customer is married.

Note that we could also answer this question using the Excel PivotTable in Figure 5.5. We can calculate these conditional probabilities by right-clicking on any numerical value in the body of the PivotTable and then selecting **Show Values As** and choosing **% of Row Total**. The modified Excel PivotTable is shown in Figure 5.6.

Figure 5.6 Using Excel PivotTable to Calculate Conditional Probabilities

	A	B	C	D
3	Count of Customer Number	Column Labels		
4	Row Labels	NO	YES	Grand Total
5	MARRIED	44.76%	55.24%	100.00%
6	SINGLE	73.89%	26.11%	100.00%
7	Grand Total	60.00%	40.00%	100.00%

PivotTable Fields

Choose fields to add to report:

Search

- ☑ Customer Number
- ☐ Age
- ☑ Marital Status
- ☐ Annual Income ($)
- ☐ Mortgage Amount ($)
- ☐ Payments Per Year
- ☐ Total Amount Paid ($)
- ☑ Default on Mortgage?

More Tables...

Drag fields between areas below:

Filters | Columns: Default on Mortgage?

Rows: Marital Status | Values: Count of Customer N...

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

By calculating the **% of Row Total**, the Excel PivotTable in Figure 5.6 shows that 55.24% of married customers defaulted on mortgages, but only 26.11% of single customers defaulted.

Independent Events

Note that in our example, $P(D) = 0.4000$, $P(D \mid M) = 0.5524$, and $P(D \mid S) = 0.2611$. So the probability that a customer defaults is influenced by whether the customer is married or single. Because $P(D \mid M) \neq P(D)$, we say that events D and M are dependent. However, if the probability of event D is not changed by the existence of event M—that is, if $P(D \mid M) = P(D)$—then we would say that events D and M are **independent events**. This is summarized for two events A and B as follows:

Independent Events

Two events A and B are independent if

$$P(A \mid B) = P(A) \tag{5.5}$$

or

$$P(B \mid A) = P(B) \tag{5.6}$$

Otherwise, the events are dependent.

Multiplication Law

The multiplication law can be used to calculate the probability of the intersection of two events. The multiplication law is based on the definition of conditional probability. Solving equations (5.3) and (5.4) for $P(A \cap B)$, we obtain the **multiplication law**.

Multiplication Law

$$P(A \cap B) = P(B)P(A \mid B) \tag{5.7}$$

or

$$P(A \cap B) = P(A)P(B \mid A) \tag{5.8}$$

To illustrate the use of the multiplication law, we will calculate the probability that a customer defaults on their mortgage and the customer is married, $P(D \cap M)$. From equation (5.7), this is calculated as $P(D \cap M) = P(M)P(D \mid M)$.

From Table 5.5 we know that $P(M) = 0.4766$, and from our previous calculations we know that the conditional probability $P(D \mid M) = 0.5524$. Therefore,

$$P(D \cap M) = P(M)P(D \mid M) = (0.4766)(0.5524) = 0.2633$$

This value matches the value shown for $P(D \cap M)$ in Table 5.5. The multiplication law is useful when we know conditional probabilities but do not know the joint probabilities.

Consider the special case in which events A and B are independent. From equations (5.5) and (5.6), $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$. Using these equations to simplify equations (5.7) and (5.8) for this special case, we obtain the following multiplication law for independent events.

Multiplication Law for Independent Events

$$P(A \cap B) = P(A)P(B) \tag{5.9}$$

To compute the probability of the intersection of two independent events, we simply multiply the probabilities of each event.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Bayes' Theorem

Revising probabilities when new information is obtained is an important aspect of probability analysis. Often, we begin the analysis with initial or **prior probability** estimates for specific events of interest. Then, from sources such as a sample survey or a product test, we obtain additional information about the events. Given this new information, we update the prior probability values by calculating revised probabilities, referred to as **posterior probabilities**. **Bayes' theorem** provides a means for making these probability calculations.

Bayes' theorem is also discussed in Chapter 17 in the context of decision analysis.

As an application of Bayes' theorem, consider a manufacturing firm that receives shipments of parts from two different suppliers. Let A_1 denote the event that a part is from supplier 1 and let A_2 denote the event that a part is from supplier 2. Currently, 65% of the parts purchased by the company are from supplier 1 and the remaining 35% are from supplier 2. Hence, if a part is selected at random, we would assign the prior probabilities $P(A_1) = 0.65$ and $P(A_2) = 0.35$.

The quality of the purchased parts varies according to their source. Historical data suggest that the quality ratings of the two suppliers are as shown in Table 5.6.

If we let G be the event that a part is good and we let B be the event that a part is bad, the information in Table 5.6 enables us to calculate the following conditional probability values:

$$P(G \mid A_1) = 0.98 \qquad P(B \mid A_1) = 0.02$$
$$P(G \mid A_2) = 0.95 \qquad P(B \mid A_2) = 0.05$$

Figure 5.7 shows a diagram that depicts the process of the firm receiving a part from one of the two suppliers and then discovering that the part is good or bad as a two-step random experiment. We see that four outcomes are possible; two correspond to the part being good and two correspond to the part being bad.

Each of the outcomes is the intersection of two events, so we can use the multiplication rule to compute the probabilities. For instance,

$$P(A_1, G) = P(A_1 \cap G) = P(A_1)P(G \mid A_1)$$

The process of computing these joint probabilities can be depicted in what is called a probability tree (see Figure 5.8). From left to right through the tree, the probabilities for each branch at step 1 are prior probabilities and the probabilities for each branch at step 2 are conditional probabilities. To find the probability of each experimental outcome, simply multiply the probabilities on the branches leading to the outcome. Each of these joint probabilities is shown in Figure 5.8 along with the known probabilities for each branch.

Now suppose that the parts from the two suppliers are used in the firm's manufacturing process and that a machine breaks down while attempting the process using a bad part. Given the information that the part is bad, what is the probability that it came from supplier 1 and what is the probability that it came from supplier 2? With the information in the probability tree (Figure 5.8), Bayes' theorem can be used to answer these questions.

For the case in which there are only two events (A_1 and A_2), Bayes' theorem can be written as follows:

Bayes' Theorem (Two-Event Case)

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \tag{5.10}$$

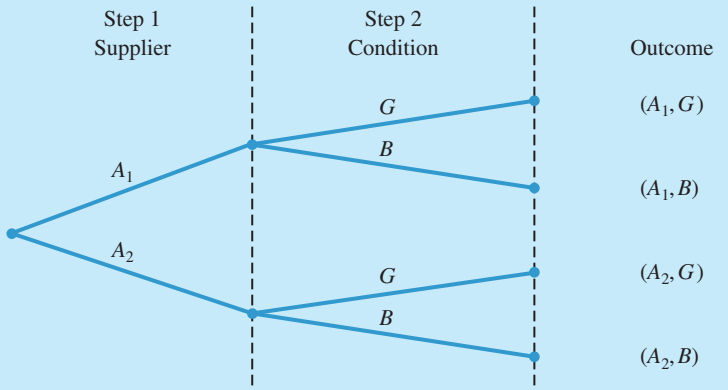
$$P(A_2 \mid B) = \frac{P(A_2)P(B \mid A_2)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \tag{5.11}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 5.6 Historical Quality Levels for Two Suppliers

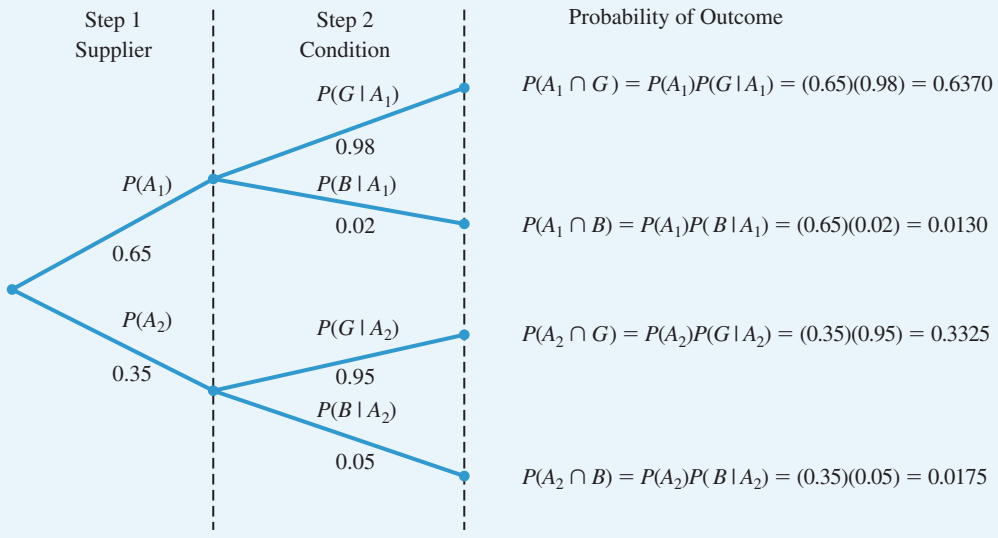
	% Good Parts	% Bad Parts
Supplier 1	98	2
Supplier 2	95	5

Figure 5.7 Diagram for Two-Supplier Example



Note: Step 1 shows that the part comes from one of two suppliers and Step 2 shows whether the part is good or bad.

Figure 5.8 Probability Tree for Two-Supplier Example



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Using equation (5.10) and the probability values provided in Figure 5.8, we have

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)}$$
$$= \frac{(0.65)(0.02)}{(0.65)(0.02) + (0.35)(0.05)} = \frac{0.0130}{0.0130 + 0.0175}$$
$$= \frac{0.0130}{0.0305} = 0.4262$$

Using equation (5.11), we find $P(A_2 \mid B)$ as

$$P(A_2 \mid B) = \frac{P(A_2)P(B \mid A_2)}{P(A_2)P(B \mid A_1) + P(A_2)P(B \mid A_2)}$$
$$= \frac{(0.35)(0.05)}{(0.65)(0.02) + (0.35)(0.05)} = \frac{0.0175}{0.0130 + 0.0175}$$
$$= \frac{0.0175}{0.0305} = 0.5738$$

Note that in this application we started with a probability of 0.65 that a part selected at random was from supplier 1. However, given information that the part is bad, the probability that the part is from supplier 1 drops to 0.4262. In fact, if the part is bad, the chance is better than 50–50 that it came from supplier 2; that is, $P(A_2 \mid B) = 0.5738$.

If the union of events is the entire sample space, the events are said to be collectively exhaustive.

Bayes' theorem is applicable when events for which we want to compute posterior probabilities are mutually exclusive and their union is the entire sample space. For the case of n mutually exclusive events $A_1, A_2, \ldots, A_n$, whose union is the entire sample space, Bayes' theorem can be used to compute any posterior probability $P(A_i \mid B)$ as shown in equation (5.12).

Bayes' Theorem

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n)} \quad \textbf{(5.12)}$$

Notes + Comments

By applying basic algebra, we can derive the multiplication law from the definition of conditional probability. For two events A and B, the probability of A given B is $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$. If we multiply both sides of this expression by $P(B)$, the $P(B)$ in the numerator and denominator on the right side of the expression will cancel and we are left with $P(A \mid B)P(B) = P(A \cap B)$, which is the multiplication law.

5.4 Random Variables

Chapter 2 introduces the concept of random variables and the use of data to describe them.

In probability terms, a **random variable** is a numerical description of the outcome of a random experiment. Because the outcome of a random experiment is not known with certainty, a random variable can be thought of as a quantity whose value is not known with certainty. A random variable can be classified as being either discrete or continuous depending on the numerical values it can assume.

Discrete Random Variables

A random variable that can take on only specified discrete values is referred to as a **discrete random variable**. Table 5.7 provides examples of discrete random variables.

Returning to our example of Lancaster Savings and Loan, we can define a random variable x to indicate whether or not a customer defaults on their mortgage. As previously

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 5.7 Examples of Discrete Random Variables

Random Experiment	Random Variable (x)	Possible Values for the Random Variable
Flip a coin	Face of coin showing	1 if heads; 0 if tails
Roll a six-sided die	Number of dots showing on top of die	1, 2, 3, 4, 5, 6
Contact five customers	Number of customers who place an order	0, 1, 2, 3, 4, 5
Operate a health care clinic for one day	Number of patients who arrive	0, 1, 2, 3, …
Offer a customer the choice of two products	Product chosen by customer	0 if none; 1 if choose product A; 2 if choose product B

stated, the values of a random variable must be numerical, so we can define random variable x such that $x = 1$ if the customer defaults on their mortgage and $x = 0$ if the customer does not default on their mortgage. An additional random variable, y, could indicate whether the customer is married or single. For instance, we can define random variable y such that $y = 1$ if the customer is married and $y = 0$ if the customer is single. Yet another random variable, z, could be defined as the number of mortgage payments per year made by the customer. For instance, a customer who makes monthly payments would make $z = 12$ payments per year, a customer who makes payments quarterly would make $z = 4$ payments per year.

Table 5.8 repeats the joint probability table for the Lancaster Savings and Loan data, but this time with the values labeled as random variables.

Continuous Random Variables

A random variable that may assume any numerical value in an interval or collection of intervals is called a **continuous random variable**. Technically, relatively few random variables are truly continuous; these include values related to time, weight, distance, and temperature. An example of a continuous random variable is $x =$ the time between consecutive incoming calls to a call center. This random variable can take on any value $x > 0$ such as $x = 1.26$ minutes, $x = 2.571$ minutes, or $x = 4.3333$ minutes. Table 5.9 provides examples of continuous random variables.

As illustrated by the final example in Table 5.9, many discrete random variables have a large number of potential outcomes and so can be effectively modeled as continuous random variables. Consider our Lancaster Savings and Loan example. We can define a random variable $x =$ total amount paid by customer over the lifetime of the mortgage. Because we typically measure financial values only to two decimal places, one could consider this a discrete random variable. However, because in any practical interval there are many possible values for this random variable, then it is usually appropriate to model the amount as a continuous random variable.

Table 5.8 Joint Probability Table for Customer Mortgage Defaults

	No Default ($x = 0$)	Default ($x = 1$)	$f(y)$
Married ($y = 1$)	0.2133	0.2633	*0.4766*
Single ($y = 0$)	0.3867	0.1367	*0.5234*
$f(x)$	*0.6000*	*0.4000*	*1.0000*

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 5.9 Examples of Continuous Random Variables

Random Experiment	Random Variable (x)	Possible Values for the Random Variable
Customer visits a web page	Time customer spends on web page in minutes	$x \geq 0$
Fill a soft drink can (max capacity = 12.1 ounces)	Number of ounces	$0 \leq x \leq 12.1$
Test a new chemical process	Temperature when the desired reaction takes place (min temperature = 150°F; max temperature = 212°F)	$150 \leq x \leq 212$
Invest $10,000 in the stock market	Value of investment after one year	$x \geq 0$

Notes + Comments

1. In this section we again use the relative frequency method to assign probabilities for the Lancaster Savings and Loan example. Technically, the concept of random variables applies only to populations; probabilities that are found using sample data are only estimates of the true probabilities. However, larger samples generate more reliable estimated probabilities, so if we have a large enough data set (as we are assuming here for the Lancaster Savings and Loan data), then we can treat the data as if they are from a population and the relative frequency method is appropriate to assign probabilities to the outcomes.
2. Random variables can be used to represent uncertain future values. Chapter 13 explains how random variables can be used in simulation models to evaluate business decisions in the presence of uncertainty.

5.5 Discrete Probability Distributions

The **probability distribution** for a random variable describes the range and relative likelihood of possible values for a random variable. For a discrete random variable x, the probability distribution is defined by a **probability mass function**, denoted by $f(x)$. The probability mass function provides the probability for each value of the random variable.

Returning to our example of mortgage defaults, consider the data shown in Table 5.3 for Lancaster Savings and Loan and the associated joint probability table in Table 5.8. From Table 5.8, we see that $f(0) = 0.6$ and $f(1) = 0.4$. Note that these values satisfy the required conditions of a discrete probability distribution that (1) $f(x) \geq 0$ and (2) $\Sigma f(x) = 1$.

We can also present probability distributions graphically. In Figure 5.9, the values of the random variable x are shown on the horizontal axis and the probability associated with these values is shown on the vertical axis.

Custom Discrete Probability Distribution

A probability distribution that is generated from observations such as that shown in Figure 5.9 is called an **empirical probability distribution**. This particular empirical probability distribution is considered a custom discrete distribution because it is discrete and the possible values of the random variable have different probabilities.

A **custom discrete probability distribution** is very useful for describing different possible scenarios that have different probabilities of occurring. The probabilities associated with each scenario can be generated using either the subjective method or the relative frequency method. Using a subjective method, probabilities are based on experience or intuition when little relevant data are available. If sufficient data exist, the relative frequency method can be used to determine probabilities. Consider the random variable describing the number of payments made per year by a randomly chosen customer. Table 5.10 presents a summary of the number of payments made per year by the 300 home mortgage

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.9 Graphical Representation of the Probability Distribution for Whether a Customer Defaults on a Mortgage

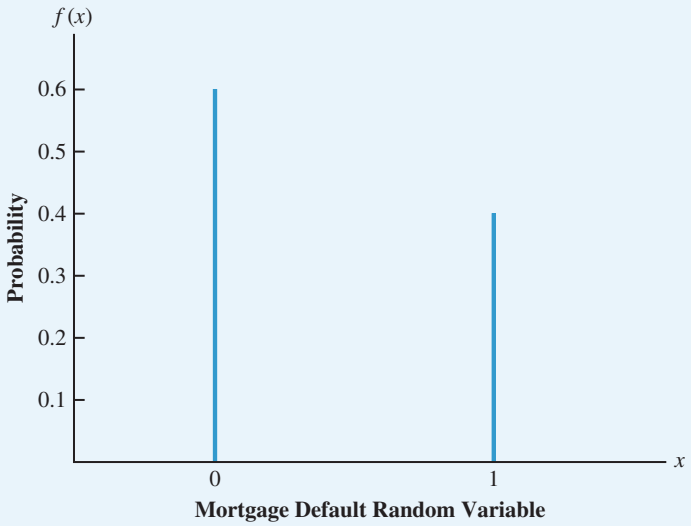


Table 5.10 Summary Table of Number of Payments Made per Year

	Number of Payments Made per Year			
	$x = 4$	$x = 12$	$x = 24$	Total
Number of observations	45	180	75	300
$f(x)$	0.15	0.60	0.25	

customers. This table shows us that 45 customers made quarterly payments ($x = 4$), 180 customers made monthly payments ($x = 12$), and 75 customers made two payments each month ($x = 24$). We can then calculate $f(4) = 45/300 = 0.15$, $f(12) = 180/300 = 0.60$, and $f(24) = 75/300 = 0.25$. In other words, the probability that a randomly selected customer makes 4 payments per year is 0.15, the probability that a randomly selected customer makes 12 payments per year is 0.60, and the probability that a randomly selected customer makes 24 payments per year is 0.25.

We can write this probability distribution as a function in the following manner:

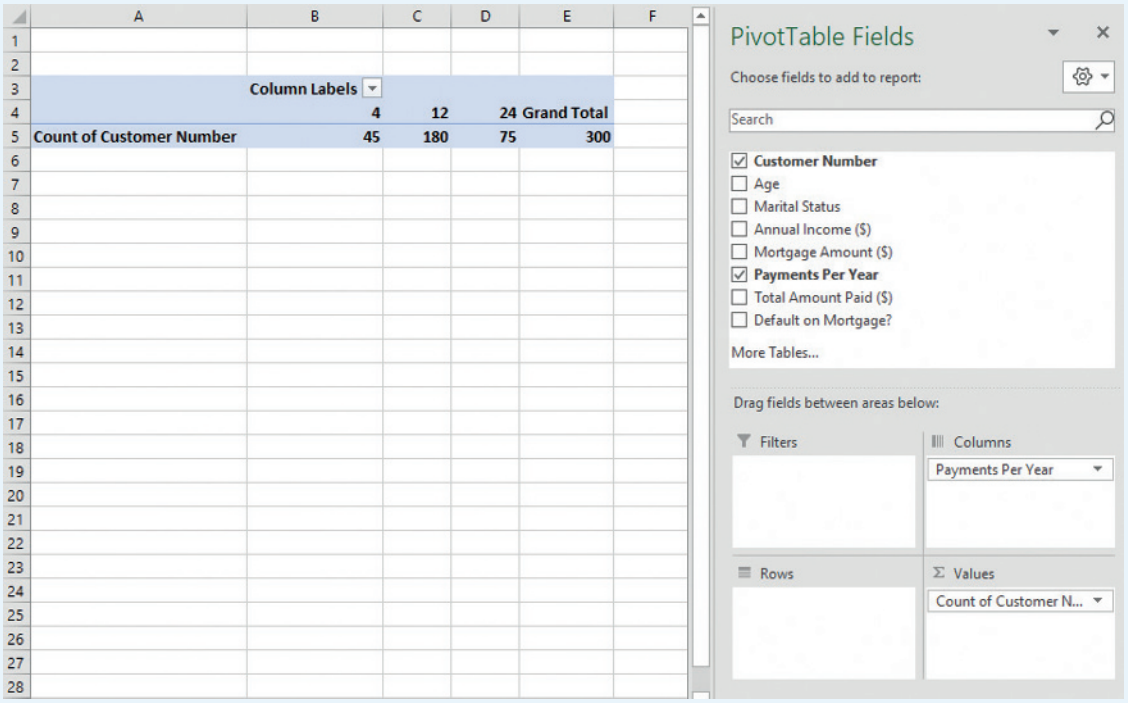
$$f(x) = \begin{cases} 0.15 \text{ if } x = 4 \\ 0.60 \text{ if } x = 12 \\ 0.25 \text{ if } x = 24 \\ 0 \quad \text{otherwise} \end{cases}$$

This probability mass function tells us in a convenient way that $f(x) = 0.15$ when $x = 4$ (the probability that the random variable $x = 4$ is 0.15); $f(x) = 0.60$ when $x = 12$ (the probability that the random variable $x = 12$ is 0.60); $f(x) = 0.25$ when $x = 24$ (the probability that the random variable $x = 24$ is 0.25); and $f(x) = 0$ when x is any other value (there is zero probability that the random variable x is some value other than 4, 12, or 24).

Note that we can also create Table 5.10 in Excel using a PivotTable as shown in Figure 5.10.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.10 Excel PivotTable for Number of Payments Made per Year



Expected Value and Variance

Chapter 2 discusses the computation of the mean of a random variable based on data.

The **expected value**, or mean, of a random variable is a measure of the central location for the random variable. It is the weighted average of the values of the random variable, where the weights are the probabilities. The formula for the expected value of a discrete random variable x follows:

Expected Value of a Discrete Random Variable

$$E(x) = \mu = \Sigma x f(x) \tag{5.13}$$

Both the notations $E(x)$ and μ are used to denote the expected value of a random variable. Equation (5.13) shows that to compute the expected value of a discrete random variable, we must multiply each value of the random variable by the corresponding probability $f(x)$ and then add the resulting products. Table 5.11 calculates the expected value of the number of payments made by a mortgage customer in a year. The sum of the entries in the $xf(x)$ column shows that the expected value is 13.8 payments per year. Therefore, if Lancaster Savings and Loan signs up a new mortgage customer, the expected number of payments per year made by this new customer is 13.8. Although, no customer will make exactly 13.8 payments per year, this value represents our expectation for the number of payments per year made by a new customer absent any other information about the new customer. Some customers will make fewer payments (4 or 12 per year), some customers will make more payments (24 per year), but 13.8 represents the expected number of payments per year based on the probabilities calculated in Table 5.10.

The SUMPRODUCT function in Excel can easily be used to calculate the expected value for a discrete random variable. This is illustrated in Figure 5.11. We can also

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 5.11 Calculation of the Expected Value for Number of Payments Made per Year by a Lancaster Savings and Loan Mortgage Customer

x	$f(x)$	$xf(x)$
4	0.15	(4)(0.15) = 0.6
12	0.60	(12)(0.60) = 7.2
24	0.25	(24)(0.25) = 6.0
		13.8 ← $E(x) = \mu = \sum xf(x)$

calculate the expected value of the random variable directly from the Lancaster Savings and Loan data using the Excel function AVERAGE, as shown in Figure 5.12. Column F contains the data on the number of payments made per year by each mortgage customer in the data set. Using the Excel formula *=AVERAGE(F2:F301)* gives us a value of 13.8 for the expected value, which is the same as the value we calculated in Table 5.11.

Note that we cannot simply use the AVERAGE function on the x values for a custom discrete random variable. If we did, this would give us a calculated value of $(4 + 12 + 24)/3 = 13.333$, which is not the correct expected value for this random variable. This is because using the AVERAGE function in this way assumes that each value of the random variable x is equally likely. But in this case, we know that $x = 12$ is much more likely than $x = 4$ or $x = 24$. Therefore, we must use equation (5.13) to calculate the expected value of a custom discrete random variable, or we can use the Excel function AVERAGE on the entire data set, as shown in Figure 5.12.

Figure 5.11 Using Excel SUMPRODUCT Function to Calculate the Expected Value for Number of Payments Made per Year by a Lancaster Savings and Loan Mortgage Customer

	A	B	C
1	*x*	*f(x)*	
2	4	0.15	
3	12	0.6	
4	24	0.25	
5			
6	**Expected Value:**	=SUMPRODUCT(A2:A4,B2:B4)	

	A	B	C
1	*x*	*f(x)*	
2	4	0.15	
3	12	0.60	
4	24	0.25	
5			
6	**Expected Value:**	13.8	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.12 Excel Calculation of the Expected Value for Number of Payments Made per Year by a Lancaster Savings and Loan Mortgage Customer

	A	B	C	D	E	F	G	H
1	Customer Number	Age	Marital Status	Annual Income ($)	Mortgage Amount ($)	Payments Per Year	Total Amount Paid ($)	Default on Mortgage?
2	1	37	SINGLE	172125.7	473402.96	24	581885.13	YES
3	2	31	SINGLE	108571.04	300468.6	12	489320.38	NO
4	3	37	MARRIED	124136.41	330664.24	24	493541.93	YES
5	4	24	MARRIED	79614.04	230222.94	24	449682.09	YES
6	5	27	SINGLE	68087.33	282203.53	12	520581.82	NO
296	295	37	MARRIED	84791.08	179676.63	24	256361.65	YES
297	296	33	MARRIED	83498.89	235907.5	12	437145.85	NO
298	297	41	SINGLE	16597.53	151972.2	4	171289.87	YES
299	298	30	SINGLE	49293.95	186043.13	12	376694.27	NO
300	299	35	SINGLE	84241.8	194417.84	12	352597.79	NO
301	300	31	MARRIED	94428.15	264175.55	24	434102.49	YES
302								
303					Expected Value:	=AVERAGE(F2:F301)		
304					Variance:	=VAR.P(F2:F301)		
305					Standard Deviation:	=STDEV.P(F2:F301)		

	A	B	C	D	E	F	G	H
1	Customer Number	Age	Marital Status	Annual Income ($)	Mortgage Amount ($)	Payments Per Year	Total Amount Paid ($)	Default on Mortgage?
2	1	37	SINGLE	172,125.70	473,402.96	24	581,885.13	YES
3	2	31	SINGLE	108,571.04	300,468.60	12	489,320.38	NO
4	3	37	MARRIED	124,136.41	330,664.24	24	493,541.93	YES
5	4	24	MARRIED	79,614.04	230,222.94	24	449,682.09	YES
6	5	27	SINGLE	68,087.33	282,203.53	12	520,581.82	NO
296	295	37	MARRIED	84,791.08	179,676.63	24	256,361.65	YES
297	296	33	MARRIED	83,498.89	235,907.50	12	437,145.85	NO
298	297	41	SINGLE	16,597.53	151,972.20	4	171,289.87	YES
299	298	30	SINGLE	49,293.95	186,043.13	12	376,694.27	NO
300	299	35	SINGLE	84,241.80	194,417.84	12	352,597.79	NO
301	300	31	MARRIED	94,428.15	264,175.55	24	434,102.49	YES
302								
303					Expected Value:	13.800		
304					Variance:	42.360		
305					Standard Deviation:	6.508		

Chapter 2 discusses the computation of the variance of a random variable based on data.

Variance is a measure of variability in the values of a random variable. It is a weighted average of the squared deviations of a random variable from its mean where the weights are the probabilities. Below we define the formula for calculating the variance of a discrete random variable.

Variance of a Discrete Random Variable

$$\text{Var}(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x) \tag{5.14}$$

As equation (5.14) shows, an essential part of the variance formula is the deviation, $x - \mu$, which measures how far a particular value of the random variable is from the expected value, or mean, μ. In computing the variance of a random variable, the deviations are squared and then weighted by the corresponding value of the probability mass function. The sum of these weighted squared deviations for all values of the random variable is referred to as the *variance*. The notations Var(x) and σ^2 are both used to denote the variance of a random variable.

The calculation of the variance of the number of payments made per year by a mortgage customer is summarized in Table 5.12. We see that the variance is 42.360. The **standard deviation**, σ, is defined as the positive square root of the variance. Thus, the standard deviation for the number of payments made per year by a mortgage customer is $\sqrt{42.360} = 6.508$.

Chapter 2 discusses the computation of the standard deviation of a random variable based on data.

The Excel function SUMPRODUCT can be used to easily calculate equation (5.14) for a custom discrete random variable. We illustrate the use of the SUMPRODUCT function to calculate variance in Figure 5.13.

We can also use Excel to find the variance directly from the data when the values in the data occur with relative frequencies that correspond to the probability distribution of the random variable. Cell F304 in Figure 5.12 shows that we use the Excel formula

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 5.12 Calculation of the Variance for Number of Payments Made per Year by a Lancaster Savings and Loan Mortgage Customer

x	$x - \mu$	$f\ x$	$(x - \mu)^2 f(x)$
4	$4 - 13.8 = -9.8$	0.15	$-9.8\ ^2\ 0.15 = 15.606$
12	$12 - 13.8 = -1.8$	0.60	$-1.8\ ^2\ 0.60 = 2.904$
24	$21 - 13.8 = 10.2$	0.25	$10.2\ ^2\ 0.25 = 24.010$
			42.360 ← $\sigma^2 = \sum(x - \mu)^2 f(x)$

Figure 5.13 Excel Calculation of the Variance for Number of Payments Made per Year by a Lancaster Savings and Loan Mortgage Customer

	A	B	C
1	x	$f(x)$	$(x-\mu)^2$
2	4	0.15	=(A2-$B$6)^2
3	12	0.6	=(A3-$B$6)^2
4	24	0.25	=(A4-$B$6)^2
5			
6	**Expected Value:**	=SUMPRODUCT(A2:A4,B2:B4)	
7	**Variance:**	=SUMPRODUCT(B2:B4,C2:C4)	
8	**Standard Deviation:**	=SQRT(B7)	

	A	B	C
1	x	$f(x)$	$(x-\mu)^2$
2	4	0.15	96.04
3	12	0.60	3.24
4	24	0.25	104.04
5			
6	**Expected Value:**	13.800	
7	**Variance:**	42.360	
8	**Standard Deviation:**	6.508	

Note that here we are using the Excel functions VAR.P *and* STDEV.P *rather than* VAR.S *and* STDEV.S. *This is because we are assuming that the sample of 300 Lancaster Savings and Loan mortgage customers is a perfect representation of the population.*

=VAR.P(F2:F301) to calculate the variance from the complete data. This formula gives us a value of 42.360, which is the same as that calculated in Table 5.12 and Figure 5.13. Similarly, we can use the formula *=STDEV.P(F2:F301)* to calculate the standard deviation of 6.508.

As with the AVERAGE function and expected value, we cannot use the Excel functions VAR.P and STDEV.P directly on the x values to calculate the variance and standard deviation of a custom discrete random variable if the x values are not equally likely to occur. Instead we must either use the formula from equation (5.14) or use the Excel functions on the entire data set as shown in Figure 5.12.

Discrete Uniform Probability Distribution

When the possible values of the probability mass function, $f(x)$, are all equal, then the probability distribution is a **discrete uniform probability distribution**. For instance, the values that result from rolling a single six-sided fair die is an example of a discrete uniform

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

distribution because the possible outcomes $y = 1$, $y = 2$, $y = 3$, $y = 4$, $y = 5$, and $y = 6$ all have the same values $f(1) = f(2) = f(3) = f(4) = f(5) = f(6) = 1/6$. The general form of the probability mass function for a discrete uniform probability distribution is as follows:

Discrete Uniform Probability Mass Function

$$f(x) = 1/n \tag{5.15}$$

where n = the number of unique values that may be assumed by the random variable.

Binomial Probability Distribution

As an example of the use of the binomial probability distribution, consider an online specialty clothing company called Martin's. Martin's commonly sends out targeted e-mails to its best customers notifying them about special discounts that are available only to the recipients of the e-mail. The e-mail contains a link that takes the customer directly to a web page for the discounted item. Before the e-mails are sent, the exact number of customers who will click on the link is unknown, but from previous data, Martin's estimates that the probability that a customer clicks on the link in the e-mail is 0.30. Martin's is interested in knowing more about the probabilities associated with one, two, three, etc. customers clicking on the link in the targeted e-mail.

The probability distribution related to the number of customers who click on the targeted e-mail link can be described using a **binomial probability distribution**. A binomial probability distribution is a discrete probability distribution that can be used to describe many situations in which a fixed number (n) of repeated identical and independent trials has two, and only two, possible outcomes. In general terms, we refer to these two possible outcomes as either a success or a failure. A success occurs with probability p in each trial and a failure occurs with probability $1 - p$ in each trial. In the Martin's example, the "trial" refers to a customer receiving the targeted e-mail. We will define a success as a customer clicking on the e-mail link ($p = 0.30$) and a failure as a customer not clicking on the link ($1 - p = 0.70$). The binomial probability distribution can then be used to calculate the probability of a given number of successes (customers who click on the e-mail link) out of a given number of independent trials (number of e-mails sent to customers). Other examples that can often be described by a binomial probability distribution include counting the number of heads resulting from flipping a coin 20 times, the number of customers who click on a particular advertisement link on web site in a day, the number of days in a month on which a particular financial stock increases in value, and the number of nondefective parts produced in a batch.

Whether or not a customer clicks on the link is an example of what is known as a Bernoulli trial—a trial in which (1) there are two possible outcomes, success or failure, and (2) the probability of success is the same every time the trial is executed. The probability distribution related to the number of successes in a set of n *independent Bernoulli trials can be described by a binomial probability distribution.*

Equation (5.16) provides the probability mass function for a binomial random variable that calculates the probability of x successes in n independent events.

Binomial Probability Mass Function

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

where

x = the number of successes
p = the probability of a success on one trial
n = the number of trials
$f(x)$ = the probability of x successes in n trials

(5.16)

and

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

n! *is read as "*n *factorial," and* n! = n × n − 1 × n − 2 × ⋯ × 2 × 1. *For example,* 4! = 4 × 3 × 2 × 1 = 24. *The Excel formula =FACT(*n*) can be used to calculate* n *factorial.*

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

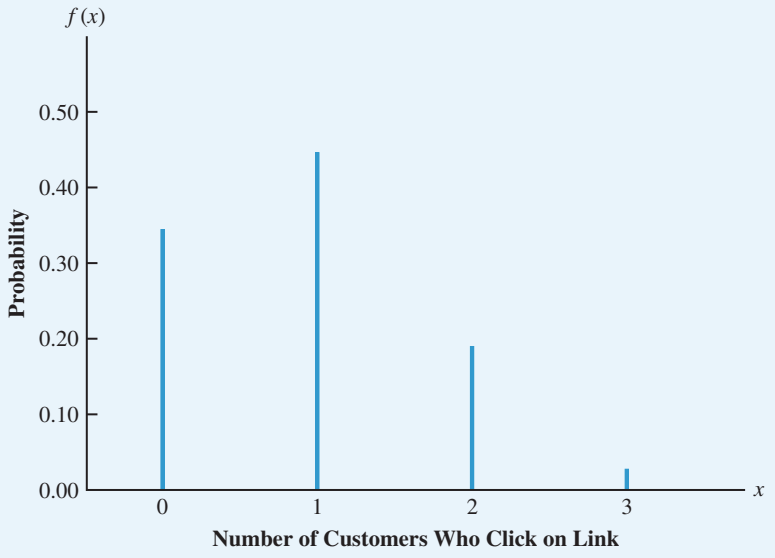
In the Martin's example, use equation (5.16) to compute the probability that out of three customers who receive the e-mail (1) no customer clicks on the link, (2) exactly one customer clicks on the link, (3) exactly two customers click on the link, and (4) all three customers click on the link. The calculations are summarized in Table 5.13, which gives the probability distribution of the number of customers who click on the targeted e-mail link. Figure 5.14 is a graph of this probability distribution. Table 5.13 and Figure 5.14 show that the highest probability is associated with exactly one customer clicking on the Martin's targeted e-mail link and the lowest probability is associated with all three customers clicking on the link.

Because the outcomes in the Martin's example are mutually exclusive, we can easily use these results to answer interesting questions about various events. For example, using the information in Table 5.13, the probability that no more than one customer clicks on the link is $P(x \leq 1) = P(x = 0) + P(x = 1) = 0.343 + 0.441 = 0.784$.

Table 5.13 Probability Distribution for the Number of Customers Who Click on the Link in the Martin's Targeted E-Mail

x	$f(x)$
0	$\frac{3!}{0!3!}(0.30)^0(0.70)^3 = 0.343$
1	$\frac{3!}{1!2!}(0.30)^1(0.70)^2 = 0.441$
2	$\frac{3!}{2!1!}(0.30)^2(0.70)^1 = 0.189$
3	$\frac{3!}{3!0!}(0.30)^3(0.70)^0 = \frac{0.027}{1.000}$

Figure 5.14 Graphical Representation of the Probability Distribution for the Number of Customers Who Click on the Link in the Martin's Targeted E-Mail



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If we consider a scenario in which 10 customers receive the targeted e-mail, the binomial probability mass function given by equation (5.16) is still applicable. If we want to find the probability that exactly 4 of the 10 customers click on the link and $p = 0.30$, then we calculate:

$$f(4) = \frac{10!}{4!6!}(0.30)^4(0.70)^6 = 0.2001$$

In Excel we can use the BINOM.DIST function to compute binomial probabilities. Figure 5.15 reproduces the Excel calculations from Table 5.13 for the Martin's problem with three customers.

The BINOM.DIST function in Excel has four input values: the first is the value of x, the second is the value of n, the third is the value of p, and the fourth is FALSE or TRUE. We choose FALSE for the fourth input if a probability mass function value $f(x)$ is desired, and TRUE if a cumulative probability is desired. The formula *=BINOM.DIST(A5,$D$1:$D$2,FALSE)* has been entered into cell B5 to compute the probability of 0 successes in three trials, $f(0)$. Figure 5.15 shows that this value is 0.343, the same as in Table 5.13.

Cells C5:C8 show the cumulative probability distribution values for this example. Note that these values are computed in Excel by entering TRUE as the fourth input in the BINOM.DIST. The cumulative probability for x using a binomial distribution is the probability of x or fewer successes out of n trials. Cell C5 computes the cumulative probability for $x = 0$, which is the same as the probability for $x = 0$ because the probability of 0 successes is the same as the probability of 0 or fewer successes. Cell C7 computes the cumulative probability for $x = 2$ using the formula *=BINOM.DIST(A7,$D$1,$D$2,TRUE)*. This value is 0.973, meaning that the probability that two or fewer customers click on the targeted e-mail link is 0.973. Note that the value 0.973 simply corresponds to $f(0) + f(1) + f(2) = 0.343 + 0.441 + 0.189 = 0.973$ because it is the probability of two or fewer customers clicking on the link, which could be zero customers, one customer, or two customers.

Figure 5.15 Excel Worksheet for Computing Binomial Probabilities of the Number of Customers Who Make a Purchase at Martin's

	A	B	C	D
1			**Number of Trials (*n*):**	3
2			**Probability of Success (*p*):**	0.3
3				
4	***x***	***f(x)***	**Cumulative Probability**	
5	0	=BINOM.DIST(A5,$D$1,$D$2,FALSE)	=BINOM.DIST(A5,$D$1,$D$2,TRUE)	
6	1	=BINOM.DIST(A6,$D$1,$D$2,FALSE)	=BINOM.DIST(A6,$D$1,$D$2,TRUE)	
7	2	=BINOM.DIST(A7,$D$1,$D$2,FALSE)	=BINOM.DIST(A7,$D$1,$D$2,TRUE)	
8	3	=BINOM.DIST(A8,$D$1,$D$2,FALSE)	=BINOM.DIST(A8,$D$1,$D$2,TRUE)	

	A	B	C	D
1			**Number of Trials (*n*):**	3
2			**Probability of Success (*p*):**	0.3
3				
4	***x***	***f(x)***	**Cumulative Probability**	
5	0	0.343	0.343	
6	1	0.441	0.784	
7	2	0.189	0.973	
8	3	0.027	1.000	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Poisson Probability Distribution

In this section, we consider a discrete random variable that is often useful in estimating the number of occurrences of an event over a specified interval of time or space. For example, the random variable of interest might be the number of patients who arrive at a health care clinic in 1 hour, the number of computer-server failures in a month, the number of repairs needed in 10 miles of highway, or the number of leaks in 100 miles of pipeline. If the following two properties are satisfied, the number of occurrences is a random variable that is described by the **Poisson probability distribution**: (1) the probability of an occurrence is the same for any two intervals (of time or space) of equal length; and (2) the occurrence or nonoccurrence in any interval (of time or space) is independent of the occurrence or nonoccurrence in any other interval.

The Poisson probability mass function is defined by equation (5.17).

Poisson Probability Mass Function

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \tag{5.17}$$

where

$f(x)$ = the probability of x occurrences in an interval
μ = expected value or mean number of occurrences in an interval
$e \approx 2.71828$

The number e is a mathematical constant that is the base of the natural logarithm. Although it is an irrational number, 2.71828 is a sufficient approximation for our purposes.

For the Poisson probability distribution, x is a discrete random variable that indicates the number of occurrences in the interval. Since there is no stated upper limit for the number of occurrences, the probability mass function $f(x)$ is applicable for values $x = 0,1,2,\ldots$ without limit. In practical applications, x will eventually become large enough so that $f(x)$ is approximately zero and the probability of any larger values of x becomes negligible.

Suppose that we are interested in the number of patients who arrive at the emergency room of a large hospital during a 15-minute period on weekday mornings. Beforehand, we do not know exactly how many patients will arrive at the emergency room in any defined interval of time, so the value of this variable is uncertain. It is important for administrators at the hospital to understand the probabilities associated with the number of arriving patients, as this information will have an impact on staffing decisions such as how many nurses and doctors to hire. It will also provide insight into possible wait times for patients to be seen once they arrive at the emergency room. If we can assume that the probability of a patient arriving is the same for any two periods of equal length during this 15-minute period and that the arrival or nonarrival of a patient in any period is independent of the arrival or nonarrival in any other period during the 15-minute period, the Poisson probability mass function is applicable. Suppose these assumptions are satisfied and an analysis of historical data shows that the average number of patients arriving during a 15-minute period of time is 10; in this case, the following probability mass function applies:

$$f(x) = \frac{10^x e^{-10}}{x!}$$

The random variable here is x = number of patients arriving at the emergency room during any 15-minute period.

If the hospital's management team wants to know the probability of exactly five arrivals during 15 minutes, we would set $x = 5$ and obtain:

$$\text{Probability of exactly 5 arrivals in 15 minutes} = f(5) = \frac{10^5 e^{-10}}{5!} = 0.0378$$

In the preceding example, the mean of the Poisson distribution is $\mu = 10$ arrivals per 15-minute period. A property of the Poisson distribution is that the mean of the distribution

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

and the variance of the distribution are *always equal.* Thus, the variance for the number of arrivals during all 15-minute periods is $\sigma^2 = 10$, and so the standard deviation is $\sigma = \sqrt{10} = 3.16$. Our illustration involves a 15-minute period, but other amounts of time can be used. Suppose we want to compute the probability of one arrival during a 3-minute period. Because 10 is the expected number of arrivals during a 15-minute period, we see that $10/15 = 2/3$ is the expected number of arrivals during a 1-minute period and that $(2/3)(3\,\text{minutes}) = 2$ is the expected number of arrivals during a 3-minute period. Thus, the probability of x arrivals during a 3-minute period with $\mu = 2$ is given by the following Poisson probability mass function:

$$f(x) = \frac{2^x e^{-2}}{x!}$$

The probability of one arrival during a 3-minute period is calculated as follows:

$$\text{Probability of exactly 1 arrival in 3 minutes} = f(1) = \frac{2^1 e^{-2}}{1!} = 0.2707$$

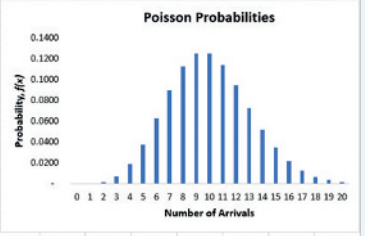
Earlier we computed the probability of five arrivals during a 15-minute period as 0.0378. Because the event of five arrivals during a 15-minute period corresponds to the same proportion as one arrival during a 3-minute period, one might expect we would get the same probability for one arrival during a 3-minute period as we do for five arrivals during a 15-minute period. However, note that the probability of one arrival during a 3-minute period is 0.2707, which is not the same. When computing a Poisson probability for a different time interval, we must first convert the mean arrival rate to the period of interest and then compute the probability.

In Excel we can use the POISSON.DIST function to compute Poisson probabilities. Figure 5.16 shows how to calculate the probabilities of patient arrivals at the emergency room if patients arrive at a mean rate of 10 per 15-minute interval.

Figure 5.16 Excel Worksheet for Computing Poisson Probabilities of the Number of Patients Arriving at the Emergency Room

	A	B	C	D
1			Mean Number of Occurrences:	10
2				
3	Number of Arrivals	Probability, f(x)	Cumulative Probability	
4	0	=POISSON.DIST(A4,$D$1,FALSE)	=POISSON.DIST(A4,$D$1,TRUE)	
5	1	=POISSON.DIST(A5,$D$1,FALSE)	=POISSON.DIST(A5,$D$1,TRUE)	
6	2	=POISSON.DIST(A6,$D$1,FALSE)	=POISSON.DIST(A6,$D$1,TRUE)	
7	3	=POISSON.DIST(A7,$D$1,FALSE)	=POISSON.DIST(A7,$D$1,TRUE)	
8	4	=POISSON.DIST(A8,$D$1,FALSE)	=POISSON.DIST(A8,$D$1,TRUE)	
9	5	=POISSON.DIST(A9,$D$1,FALSE)	=POISSON.DIST(A9,$D$1,TRUE)	
10	6	=POISSON.DIST(A10,$D$1,FALSE)	=POISSON.DIST(A10,$D$1,TRUE)	
11	7	=POISSON.DIST(A11,$D$1,FALSE)	=POISSON.DIST(A11,$D$1,TRUE)	
12	8	=POISSON.DIST(A12,$D$1,FALSE)	=POISSON.DIST(A12,$D$1,TRUE)	
13	9	=POISSON.DIST(A13,$D$1,FALSE)	=POISSON.DIST(A13,$D$1,TRUE)	
14	10	=POISSON.DIST(A14,$D$1,FALSE)	=POISSON.DIST(A14,$D$1,TRUE)	
15	11	=POISSON.DIST(A15,$D$1,FALSE)	=POISSON.DIST(A15,$D$1,TRUE)	
16	12	=POISSON.DIST(A16,$D$1,FALSE)	=POISSON.DIST(A16,$D$1,TRUE)	
17	13	=POISSON.DIST(A17,$D$1,FALSE)	=POISSON.DIST(A17,$D$1,TRUE)	
18	14	=POISSON.DIST(A18,$D$1,FALSE)	=POISSON.DIST(A18,$D$1,TRUE)	
19	15	=POISSON.DIST(A19,$D$1,FALSE)	=POISSON.DIST(A19,$D$1,TRUE)	
20	16	=POISSON.DIST(A20,$D$1,FALSE)	=POISSON.DIST(A20,$D$1,TRUE)	
21	17	=POISSON.DIST(A21,$D$1,FALSE)	=POISSON.DIST(A21,$D$1,TRUE)	
22	18	=POISSON.DIST(A22,$D$1,FALSE)	=POISSON.DIST(A22,$D$1,TRUE)	
23	19	=POISSON.DIST(A23,$D$1,FALSE)	=POISSON.DIST(A23,$D$1,TRUE)	
24	20	=POISSON.DIST(A24,$D$1,FALSE)	=POISSON.DIST(A24,$D$1,TRUE)	

	A	B	C	D
1			Mean Number of Occurrences:	10
2				
3	Number of Arrivals	Probability, f(x)	Cumulative Probability	
4	0	0.0000	0.0000	
5	1	0.0005	0.0005	
6	2	0.0023	0.0028	
7	3	0.0076	0.0103	
8	4	0.0189	0.0293	
9	5	0.0378	0.0671	
10	6	0.0631	0.1301	
11	7	0.0901	0.2202	
12	8	0.1126	0.3328	
13	9	0.1251	0.4579	
14	10	0.1251	0.5830	
15	11	0.1137	0.6968	
16	12	0.0948	0.7916	
17	13	0.0729	0.8645	
18	14	0.0521	0.9165	
19	15	0.0347	0.9513	
20	16	0.0217	0.9730	
21	17	0.0128	0.9857	
22	18	0.0071	0.9928	
23	19	0.0037	0.9965	
24	20	0.0019	0.9984	



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The POISSON.DIST function in Excel has three input values: the first is the value of x, the second is the mean of the Poisson distribution, and the third is FALSE or TRUE. We choose FALSE for the third input if a probability mass function value $f(x)$ is desired, and TRUE if a cumulative probability is desired. The formula *=POISSON.DIST(A4,$D$1,FALSE)* has been entered into cell B4 to compute the probability of 0 occurrences, $f(0)$. Figure 5.16 shows that this value (to four decimal places) is 0.0000, which means that it is highly unlikely (probability near 0) that we will have 0 patient arrivals during a 15-minute interval. The value in cell B12 shows that the probability that there will be exactly eight arrivals during a 15-minute interval is 0.1126.

The cumulative probability for x using a Poisson distribution is the probability of x or fewer occurrences during the interval. Cell C4 computes the cumulative probability for $x = 0$, which is the same as the probability for $x = 0$ because the probability of 0 occurrences is the same as the probability of 0 or fewer occurrences. Cell C12 computes the cumulative probability for $x = 8$ using the formula *=POISSON.DIST(A12,$D$1,TRUE)*. This value is 0.3328, meaning that the probability that eight or fewer patients arrive during a 15-minute interval is 0.3328. This value corresponds to

$$f(0) + f(1) + f(2) + \cdots + f(7) + f(8) = 0.0000 + 0.0005 + 0.0023 + \cdots + 0.0901 + 0.1126 = 0.3328$$

Let us illustrate an application not involving time intervals in which the Poisson distribution is useful. Suppose we want to determine the occurrence of major defects in a highway one month after it has been resurfaced. We assume that the probability of a defect is the same for any two highway intervals of equal length and that the occurrence or nonoccurrence of a defect in any one interval is independent of the occurrence or nonoccurrence of a defect in any other interval. Hence, the Poisson distribution can be applied.

Suppose we learn that major defects one month after resurfacing occur at the average rate of two per mile. Let us find the probability of no major defects in a particular 3-mile section of the highway. Because we are interested in an interval with a length of 3 miles, $\mu = (2\,\text{defects/mile})(3\,\text{miles}) = 6$ defects represents the expected number of major defects over the 3-mile section of highway. Using equation (5.17), the probability of no major defects is $f(0) = \dfrac{6^0 e^{-6}}{0!} = 0.0025$. Thus, it is unlikely that no major defects will occur in the 3-mile section. In fact, this example indicates a $1 - 0.0025 = 0.9975$ probability of at least one major defect in the 3-mile highway section.

Notes + Comments

1. If sample data are used to estimate the probabilities of a custom discrete distribution, equation (5.13) yields the sample mean $\bar{x}$ rather than the population mean μ. However, as the sample size increases, the sample generally becomes more representative of the population and the sample mean $\bar{x}$ converges to the population mean μ. In this chapter we have assumed that the sample of 300 Lancaster Savings and Loan mortgage customers is sufficiently large to be representative of the population of mortgage customers at Lancaster Savings and Loan.
2. We can use the Excel function AVERAGE only to compute the expected value of a custom discrete random variable when the values in the data occur with relative frequencies that correspond to the probability distribution of the random variable. If this assumption is not satisfied, then the estimate of the expected value with the AVERAGE function will be inaccurate. In practice, this assumption is satisfied with an increasing degree of accuracy as the size of the sample is increased. Otherwise, we must use equation (5.13) to calculate the expected value for a custom discrete random variable.
3. If sample data are used to estimate the probabilities for a custom discrete distribution, equation (5.14) yields the sample variance s^2 rather than the population variance σ^2. However, as the sample size increases the sample generally becomes more representative of the population and the sample variance s^2 converges to the population variance σ^2.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

5.6 Continuous Probability Distributions

In the preceding section we discussed discrete random variables and their probability distributions. In this section we consider continuous random variables. Specifically, we discuss some of the more useful continuous probability distributions for analytics models: the uniform, the triangular, the normal, and the exponential.

A fundamental difference separates discrete and continuous random variables in terms of how probabilities are computed. For a discrete random variable, the probability mass function $f(x)$ provides the probability that the random variable assumes a particular value. With continuous random variables, the counterpart of the probability mass function is the **probability density function**, also denoted by $f(x)$. The difference is that the probability density function does not directly provide probabilities. However, the area under the graph of $f(x)$ corresponding to a given interval does provide the probability that the continuous random variable x assumes a value in that interval. So when we compute probabilities for continuous random variables, we are computing the probability that the random variable assumes any value in an interval. Because the area under the graph of $f(x)$ at any particular point is zero, one of the implications of the definition of probability for continuous random variables is that the probability of any particular value of the random variable is zero.

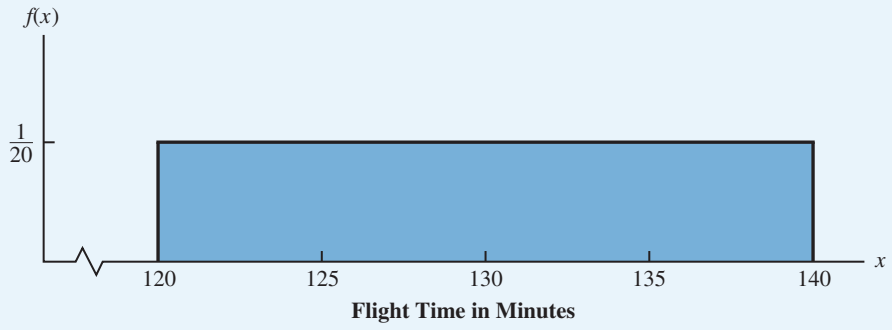
Uniform Probability Distribution

Consider the random variable x representing the flight time of an airplane traveling from Chicago to New York. The exact flight time from Chicago to New York is uncertain because it can be affected by weather (headwinds or storms), flight traffic patterns, and other factors that cannot be known with certainty. It is important to characterize the uncertainty associated with the flight time because this can have an impact on connecting flights and how we construct our overall flight schedule. Suppose the flight time can be any value in the interval from 120 minutes to 140 minutes. Because the random variable x can assume any value in that interval, x is a continuous rather than a discrete random variable. Let us assume that sufficient actual flight data are available to conclude that the probability of a flight time within any interval of a given length is the same as the probability of a flight time within any other interval of the same length that is contained in the larger interval from 120 to 140 minutes. With every interval of a given length being equally likely, the random variable x is said to have a **uniform probability distribution**. The probability density function, which defines the uniform distribution for the flight-time random variable, is:

$$f(x) = \begin{cases} 1/20 & \text{for } 120 \le x \le 140 \\ 0 & \text{elsewhere} \end{cases}$$

Figure 5.17 shows a graph of this probability density function.

Figure 5.17 Uniform Probability Distribution for Flight Time



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In general, the uniform probability density function for a random variable x is defined by the following formula:

Uniform Probability Density Function

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{elsewhere} \end{cases} \qquad \textbf{(5.18)}$$

For the flight-time random variable, $a = 120$ and $b = 140$.

For a continuous random variable, we consider probability only in terms of the likelihood that a random variable assumes a value within a specified interval. In the flight time example, an acceptable probability question is: What is the probability that the flight time is between 120 and 130 minutes? That is, what is $P(120 \le x \le 130)$?

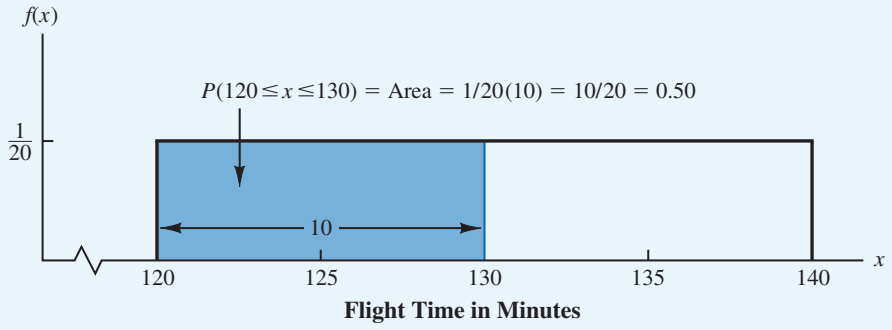
To answer this question, consider the area under the graph of $f(x)$ in the interval from 120 to 130 (see Figure 5.18). The area is rectangular, and the area of a rectangle is simply the width multiplied by the height. With the width of the interval equal to $130 - 120 = 10$ and the height equal to the value of the probability density function $f(x) = 1/20$, we have area = width × height = $10(1/20) = 10/20 = 0.50$.

The area under the graph of $f(x)$ and probability are identical for all continuous random variables. Once a probability density function $f(x)$ is identified, the probability that x takes a value between some lower value x_1 and some higher value x_2 can be found by computing the area under the graph of $f(x)$ over the interval from x_1 to x_2.

Given the uniform distribution for flight time and using the interpretation of area as probability, we can answer any number of probability questions about flight times. For example:

- What is the probability of a flight time between 128 and 136 minutes? The width of the interval is $136 - 128 = 8$. With the uniform height of $f(x) = 1/20$, we see that $P(128 \le x \le 136) = 8(1/20) = 0.40$.
- What is the probability of a flight time between 118 and 123 minutes? The width of the interval is $123 - 118 = 5$, but the height is $f(x) = 0$ for $118 \le x < 120$ and $f(x) = 1/20$ for $120 \le x \le 123$, so we have that $P(118 \le x \le 123) = P(118 \le x < 120) + P(120 \le x \le 123) = 2(0) + 3(1/20) = 0.15$.

Figure 5.18 The Area Under the Graph Provides the Probability of a Flight Time Between 120 and 130 Minutes



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note that $P(120 \le x \le 140) = 20(1/20) = 1$; that is, the total area under the graph of $f(x)$ is equal to 1. This property holds for all continuous probability distributions and is the analog of the condition that the sum of the probabilities must equal 1 for a discrete probability mass function.

Note also that because we know that the height of the graph of $f(x)$ for a uniform distribution is $\dfrac{1}{b-a}$ for $a \le x \le b$, then the area under the graph of $f(x)$ for a uniform distribution evaluated from a to a point x_0 when $a \le x_0 \le b$ is width $\times$ height $= (x_0 - a) \times (b - a)$. This value provides the cumulative probability of obtaining a value for a uniform random variable of less than or equal to some specific value denoted by x_0 and the formula is given in equation (5.19).

Uniform Distribution: Cumulative Probabilities

$$P(x \le x_0) = \frac{x_0 - a}{b - a} \text{ for } a \le x_0 \le b \tag{5.19}$$

The calculation of the expected value and variance for a continuous random variable is analogous to that for a discrete random variable. However, because the computational procedure involves integral calculus, we do not show the calculations here.

For the uniform continuous probability distribution introduced in this section, the formulas for the expected value and variance are as follows:

$$E(x) = \frac{a + b}{2}$$

$$\text{Var}(x) = \frac{(b - a)^2}{12}$$

In these formulas, a is the minimum value and b is the maximum value that the random variable may assume.

Applying these formulas to the uniform distribution for flight times from Chicago to New York, we obtain

$$E(x) = \frac{(120 + 140)}{2} = 130$$

$$\text{Var}(x) = \frac{(140 - 120)^2}{12} = 33.3333$$

The standard deviation of flight times can be found by taking the square root of the variance. Thus, for flight times from Chicago to New York, $\sigma = \sqrt{33.3333} = 5.7735$ minutes.

Triangular Probability Distribution

The triangular probability distribution is useful when only subjective probability estimates are available. There are many situations for which we do not have sufficient data and only subjective estimates of possible values are available. In the **triangular probability distribution**, we need only to specify the minimum possible value a, the maximum possible value b, and the most likely value (or mode) of the distribution m. If these values can be knowledgeably estimated for a continuous random variable by a subject-matter expert, then as an approximation of the actual probability density function, we can assume that the triangular distribution applies.

Consider a situation in which a project manager is attempting to estimate the time that will be required to complete an initial assessment of the capital project of constructing a new corporate headquarters. The assessment process includes completing environmental-impact studies, procuring the required permits, and lining up all the contractors and

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

subcontractors needed to complete the project. There is considerable uncertainty regarding the duration of these tasks, and generally little or no historical data are available to help estimate the probability distribution for the time required for this assessment process.

Suppose that we are able to discuss this project with several subject-matter experts who have worked on similar projects. From these expert opinions and our own experience, we estimate that the minimum required time for the initial assessment phase is six months and that the worst-case estimate is that this phase could require 24 months if we are delayed in the permit process or if the results from the environmental-impact studies require additional action. While a time of six months represents a best case and 24 months a worst case, the consensus is that the most likely amount of time required for the initial assessment phase of the project is 12 months. From these estimates, we can use a triangular distribution as an approximation for the probability density function for the time required for the initial assessment phase of constructing a new corporate headquarters.

Figure 5.19 shows the probability density function for this triangular distribution. Note that the probability density function is a triangular shape.

The general form of the triangular probability density function is as follows:

Triangular Probability Density Function

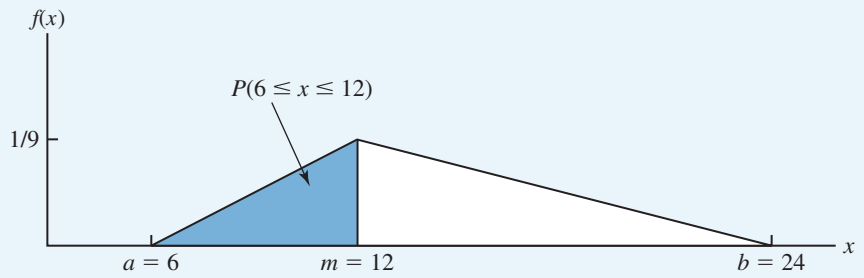
$$f(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(m-a)} & \text{for } a \le x \le m \\ \dfrac{2(b-x)}{(b-a)(b-m)} & \text{for } m < x \le b \end{cases} \tag{5.20}$$

where

a = minimum value
b = maximum value
m = mode

In the example of the time required to complete the initial assessment phase of constructing a new corporate headquarters, the minimum value a is six months, the maximum value b is 24 months, and the mode m is 12 months. As with the explanation given for the uniform distribution above, we can calculate probabilities by using the area under the graph of $f(x)$. We can calculate the probability that the time required is less than 12 months by finding the area under the graph of $f(x)$ from $x = 6$ to $x = 12$ as shown in Figure 5.19.

Figure 5.19 Triangular Probability Distribution for Time Required for Initial Assessment of Corporate Headquarters Construction



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The geometry required to find this area for any given value is slightly more complex than that required to find the area for a uniform distribution, but the resulting formula for a triangular distribution is relatively simple:

Triangular Distribution: Cumulative Probabilities

$$P(x \le x_0) = \begin{cases} \dfrac{(x_0 - a)^2}{(b - a)(m - a)} & \text{for } a \le x_0 \le m \\ 1 - \dfrac{(b - x_0)^2}{(b - a)(b - m)} & \text{for } m < x_0 \le b \end{cases} \tag{5.21}$$

Equation (5.21) provides the cumulative probability of obtaining a value for a triangular random variable of less than or equal to some specific value denoted by x_0.

To calculate $P(x \le 12)$ we use equation (5.20) with $a = 6$, $b = 24$, $m = 12$, and $x_0 = 12$.

$$P(x \le 12) = \frac{(12 - 6)^2}{(24 - 6)(12 - 6)} = 0.3333$$

Thus, the probability that the assessment phase of the project requires less than 12 months is 0.3333. We can also calculate the probability that the project requires more than 10 months, but less than or equal to 18 months by subtracting $P(x \le 10)$ from $P(x \le 18)$. This is shown graphically in Figure 5.20. The calculations are as follows:

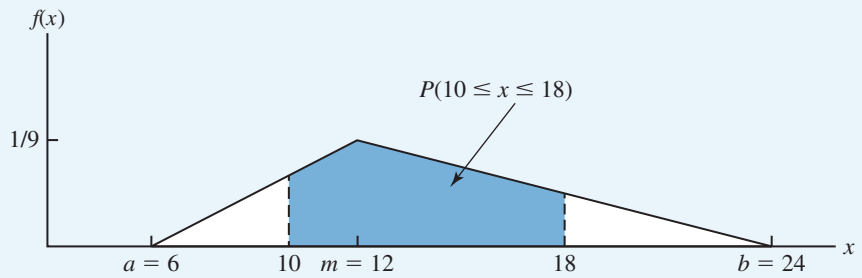
$$P(x \le 18) - P(x \le 10) = \left[1 - \frac{(24 - 18)^2}{(24 - 6)(24 - 12)}\right] - \left[\frac{(10 - 6)^2}{(24 - 6)(10 - 6)}\right] = 0.6111$$

Thus, the probability that the assessment phase of the project requires at least 10 months but less than 18 months is 0.6111.

Normal Probability Distribution

One of the most useful probability distributions for describing a continuous random variable is the **normal probability distribution**. The normal distribution has been used in a wide variety of practical applications in which the random variables are heights and weights of people, test scores, scientific measurements, amounts of rainfall, and other similar values. It is also widely used in business applications to describe uncertain quantities such as demand for products, the rate of return for stocks and bonds, and the time it takes to manufacture a part or complete many types of service-oriented activities such as medical surgeries and consulting engagements.

Figure 5.20 Triangular Distribution to Determine $P(10 \le x \le 18) = P(x \le 18) - P(x \le 10)$



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The form, or shape, of the normal distribution is illustrated by the bell-shaped normal curve in Figure 5.21.

The probability density function that defines the bell-shaped curve of the normal distribution follows.

Normal Probability Density Function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{5.22}$$

where

μ = mean
σ = standard deviation
$\pi \approx 3.14159$
$e \approx 2.71828$

Although π and e are irrational numbers, 3.14159 and 2.71828, respectively, are sufficient approximations for our purposes.

We make several observations about the characteristics of the normal distribution.

1. The entire family of normal distributions is differentiated by two parameters: the mean μ and the standard deviation σ. The mean and standard deviation are often referred to as the location and shape parameters of the normal distribution, respectively.
2. The highest point on the normal curve is at the mean, which is also the median and mode of the distribution.
3. The mean of the distribution can be any numerical value: negative, zero, or positive. Three normal distributions with the same standard deviation but three different means (−10, 0, and 20) are shown in Figure 5.22.

Figure 5.21 Bell-Shaped Curve for the Normal Distribution

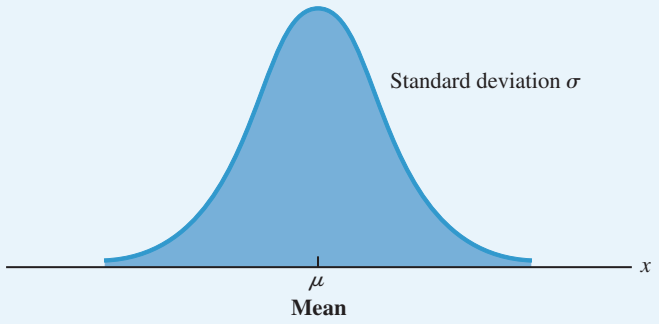
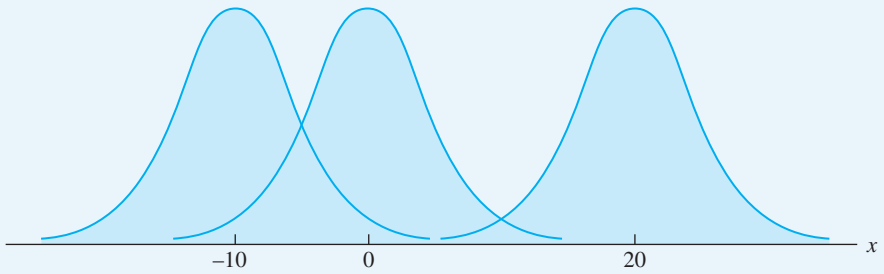


Figure 5.22 Three Normal Distributions with the Same Standard Deviation but Different Means ($\mu = -10$, $\mu = 0$, $\mu = 20$)



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4. The normal distribution is symmetric, with the shape of the normal curve to the left of the mean a mirror image of the shape of the normal curve to the right of the mean.
5. The tails of the normal curve extend to infinity in both directions and theoretically never touch the horizontal axis. Because it is symmetric, the normal distribution is not skewed; its skewness measure is zero.
6. The standard deviation determines how flat and wide the normal curve is. Larger values of the standard deviation result in wider, flatter curves, showing more variability in the data. More variability corresponds to greater uncertainty. Two normal distributions with the same mean but with different standard deviations are shown in Figure 5.23.
7. Probabilities for the normal random variable are given by areas under the normal curve. The total area under the curve for the normal distribution is 1. Because the distribution is symmetric, the area under the curve to the left of the mean is 0.50 and the area under the curve to the right of the mean is 0.50.
8. The percentages of values in some commonly used intervals are as follows:
   a. 68.3% of the values of a normal random variable are within plus or minus one standard deviation of its mean.
   b. 95.4% of the values of a normal random variable are within plus or minus two standard deviations of its mean.
   c. 99.7% of the values of a normal random variable are within plus or minus three standard deviations of its mean.

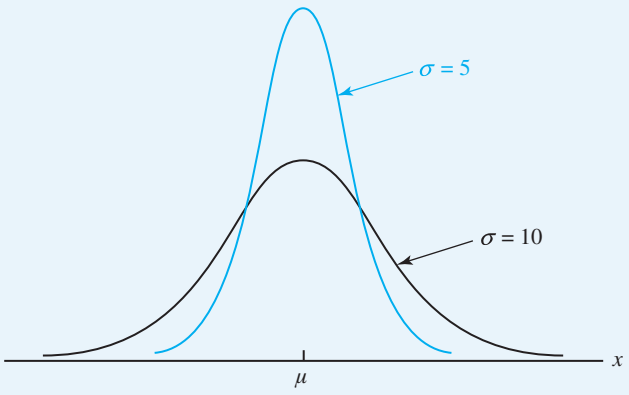
These percentages are the basis for the empirical rule discussed in Section 2.7.

Figure 5.24 shows properties (a), (b), and (c) graphically.

We turn now to an application of the normal probability distribution. Suppose Grear Aircraft Engines sells aircraft engines to commercial airlines. Grear is offering a new performance-based sales contract in which Grear will guarantee that its engines will provide a certain amount of lifetime flight hours subject to the airline purchasing a preventive-maintenance service plan that is also provided by Grear. Grear believes that this performance-based contract will lead to additional sales as well as additional income from providing the associated preventive maintenance and servicing.

From extensive flight testing and computer simulations, Grear's engineering group has estimated that if their engines receive proper parts replacement and preventive maintenance, the mean lifetime flight hours achieved is normally distributed with a mean $\mu = 36{,}500$ hours and standard deviation $\sigma = 5{,}000$ hours. Grear would like to know what percentage of its aircraft engines will be expected to last more than 40,000 hours. In other words, what is the probability that the aircraft lifetime flight hours x will exceed 40,000? This question can be answered by finding the area of the darkly shaded region in Figure 5.25.

Figure 5.23 Two Normal Distributions with the Same Mean but Different Standard Deviations ($\sigma = 5$, $\sigma = 10$)



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.24 Areas Under the Curve for Any Normal Distribution

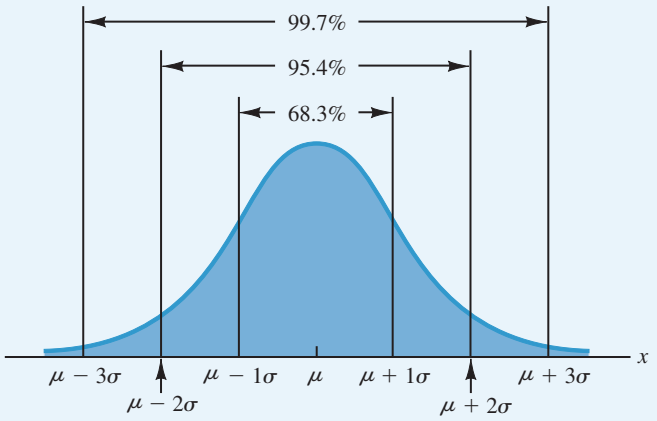
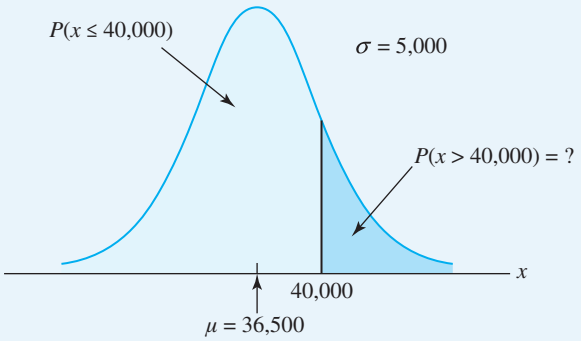


Figure 5.25 Grear Aircraft Engines Lifetime Flight Hours Distribution



The Excel function NORM.DIST can be used to compute the area under the curve for a normal probability distribution. The NORM.DIST function has four input values. The first is the value of interest corresponding to the probability you want to calculate, the second is the mean of the normal distribution, the third is the standard deviation of the normal distribution, and the fourth is TRUE or FALSE. We enter TRUE for the fourth input if we want the cumulative distribution function and FALSE if we want the probability density function.

Figure 5.26 shows how we can answer the question of interest for Grear using Excel—in cell B5, we use the formula *=NORM.DIST(40000, $B$1, $B$2, TRUE)*. Cell B1 contains the mean of the normal distribution and cell B2 contains the standard deviation. Because we want to know the area under the curve, we want the cumulative distribution function, so we use TRUE as the fourth input value in the formula. This formula provides a value of 0.7580 in cell B5. But note that this corresponds to $P(x \leq 40{,}000) = 0.7580$. In other words, this gives us the area under the curve to the left of $x = 40{,}000$ in Figure 5.25, and we are interested in the area under the curve to the right of $x = 40{,}000$. To find this value, we simply use $1 - 0.7580 = 0.2420$ (cell B6). Thus, 0.2420 is the probability that x will exceed 40,000 hours. We can conclude that about 24.2% of aircraft engines will exceed 40,000 lifetime flight hours.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.26 Excel Calculations for Grear Aircraft Engines Example

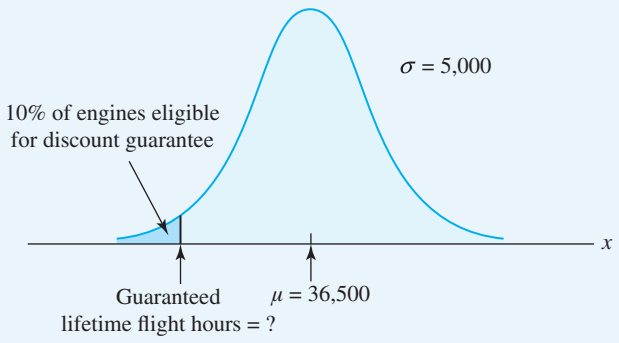
	A	B	C
1	Mean:	36500	
2	Standard Deviation:	5000	
3			
4			
5	$P(x \le 40{,}000) =$	=NORM.DIST(40000, $B$1, $B$2, TRUE)	
6	$P(x > 40{,}000) = 1 - P(x \le 40{,}000) =$	=1 - B5	
7			
8	Guarantee on lifetime flight hours for 10% of engines eligible for discount:	=NORM.INV(0.1, $B$1, $B$2)	

	A	B	C
1	Mean:	36,500	
2	Standard Deviation:	5,000	
3			
4			
5	$P(x \le 40{,}000) =$	0.7580	
6	$P(x > 40{,}000) = 1 - P(x \le 40{,}000) =$	0.2420	
7			
8	Guarantee on lifetime flight hours for 10% of engines eligible for discount:	30,092.24	

Let us now assume that Grear is considering a guarantee that will provide a discount on a replacement aircraft engine if the original engine does not meet the lifetime-flight-hour guarantee. How many lifetime flight hours should Grear guarantee if Grear wants no more than 10% of aircraft engines to be eligible for the discount guarantee? This question is interpreted graphically in Figure 5.27.

According to Figure 5.27, the area under the curve to the left of the unknown guarantee on lifetime flight hours must be 0.10. To find the appropriate value using Excel, we use the function NORM.INV. The NORM.INV function has three input values. The first is the probability of interest, the second is mean of the normal distribution, and the third is the standard deviation of the normal distribution. Figure 5.26 shows how we can use Excel to answer Grear's question about a guarantee on lifetime flight hours. In cell B8 we use the formula *=NORM.INV(0.10, $B$1, $B$2)*, where the mean of the normal distribution is

Figure 5.27 Grear's Discount Guarantee



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

With the guarantee set at 30,000 hours, the actual percentage eligible for the guarantee will be =NORM.DIST(*30000, 36500, 5000,* TRUE) = *0.0968, or 9.68%*

contained in cell B1 and the standard deviation in cell B2. This provides a value of 30,092.24. Thus, a guarantee of 30,092 hours will meet the requirement that approximately 10% of the aircraft engines will be eligible for the guarantee. This information could be used by Grear's analytics team to suggest a lifetime flight hours guarantee of 30,000 hours.

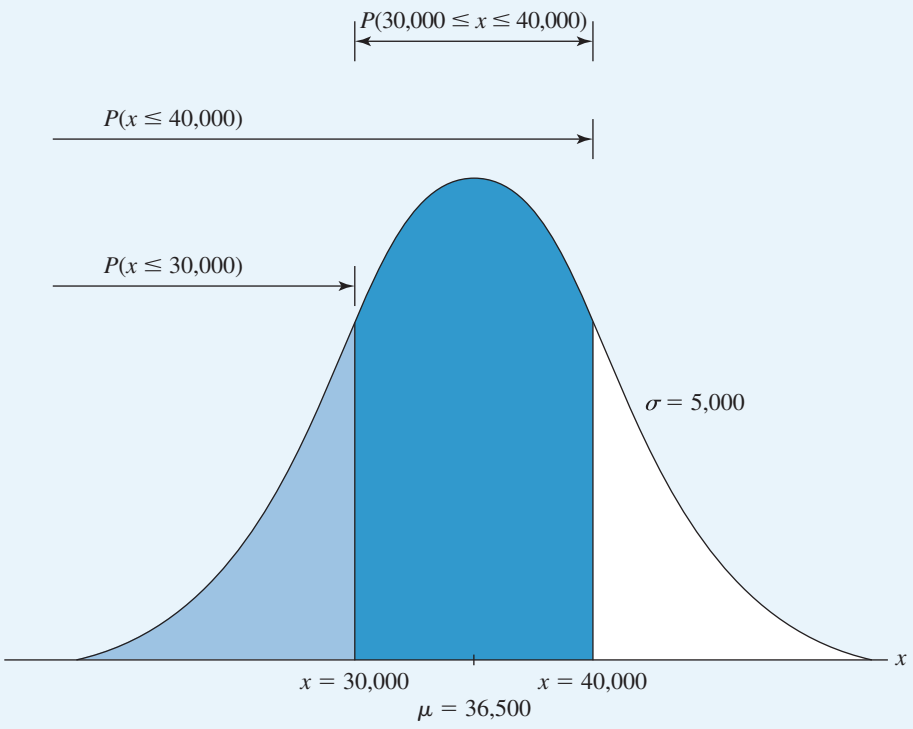
Perhaps Grear is also interested in knowing the probability that an engine will have a lifetime of flight hours greater than 30,000 hours but less than 40,000 hours. To calculate this probability, we can restate this question as follows. What is $P(30,000 \leq x \leq 40,000)$? Figure 5.28 shows the area under the curve needed to answer this question. The area that corresponds to $P(30,000 \leq x \leq 40,000)$ can be found by subtracting the area corresponding to $P(x \leq 30,000)$ from the area corresponding to $P(x \leq 40,000)$. In other words, $P(30,000 \leq x \leq 40,000) = P(x \leq 40,000) - P(x \leq 30,000)$. Figure 5.29 shows how we can find the value for $P(30,000 \leq x \leq 40,000)$ using Excel. We calculate $P(x \leq 40,000)$ in cell B5 and $P(x \leq 30,000)$ in cell B6 using the NORM.DIST function. We then calculate $P(30,000 \leq x \leq 40,000)$ in cell B8 by subtracting the value in cell B6 from the value in cell B5. This tells us that $P(30,000 \leq x \leq 40,000) = 0.7580 - 0.0968 = 0.6612$. In other words, the probability that the lifetime flight hours for an aircraft engine will be between 30,000 hours and 40,000 hours is 0.6612.

Note that we can calculate P(30,000 ≤ x ≤ 40,000) *in a single cell using the formula* =NORM.DIST(*40000, $B$1, $B$2,* TRUE) – NORM.DIST(*30000, $B$1, $B$2,* TRUE).

Exponential Probability Distribution

The **exponential probability distribution** may be used for random variables such as the time between patient arrivals at an emergency room, the distance between major defects in a highway, and the time until a borrower defaults on a loan in certain credit-risk models. The exponential probability density function is as follows:

Figure 5.28 Graph Showing the Area Under the Curve Corresponding to $P(30,000 \leq x \leq 40,000)$ in the Grear Aircraft Engines Example



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.29 Using Excel to Find $P(30{,}000 \leq x \leq 40{,}000)$ in the Grear Aircraft Engines Example

	A	B	C
1	Mean:	36500	
2	Standard Deviation:	5000	
3			
4			
5	P(x ≤ 40,000) =	=NORM.DIST(40000, $B$1, $B$2, TRUE)	
6	P(x ≤ 30,000) =	=NORM.DIST(30000, $B$1, $B$2, TRUE)	
7			
8	P(30,000 ≤ x ≤ 40,000) = P(x ≤ 40,000) - P(x ≤ 30,000) =	=B5 - B6	

	A	B	C
1	Mean:	36,500	
2	Standard Deviation:	5,000	
3			
4			
5	P(x ≤ 40,000) =	0.7580	
6	P(x ≤ 30,000) =	0.0968	
7			
8	P(30,000 ≤ x ≤ 40,000) = P(x ≤ 40,000) - P(x ≤ 30,000) =	0.6612	

Exponential Probability Density Function

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \quad \text{for } x \geq 0 \tag{5.23}$$

where

μ = expected value or mean
e = 2.71828

As an example, suppose that x represents the time between business loan defaults for a particular lending agency. If the mean, or average, time between loan defaults is 15 months ($\mu = 15$), the appropriate density function for x is

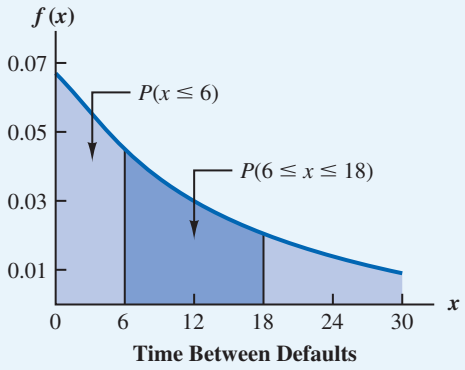
$$f(x) = \frac{1}{15} e^{-x/15}$$

Figure 5.30 is the graph of this probability density function.

As with any continuous probability distribution, the area under the curve corresponding to an interval provides the probability that the random variable assumes a value in that interval. In the time between loan defaults example, the probability that the time between two defaults is six months or less, $P(x \leq 6)$, is defined to be the area under the curve in Figure 5.30 from $x = 0$ to $x = 6$. Similarly, the probability that the time between defaults will be 18 months or less, $P(x \leq 18)$, is the area under the curve from $x = 0$ to $x = 18$. Note also that the probability that the time between defaults will be between 6 months and 18 months, $P(6 \leq x \leq 18)$, is given by the area under the curve from $x = 6$ to $x = 18$.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.30 Exponential Distribution for the Time Between Business Loan Defaults Example



To compute exponential probabilities such as those just described, we use the following formula, which provides the cumulative probability of obtaining a value for the exponential random variable of less than or equal to some specific value denoted by x_0.

Exponential Distribution: Cumulative Probabilities

$$P(x \leq x_0) = 1 - e^{-x_0/\mu} \tag{5.24}$$

For the time between defaults example, x = time between business loan defaults in months and $\mu = 15$ months. Using equation (5.24),

$$P(x \leq x_0) = 1 - e^{-x_0/15}$$

Hence, the probability that the time between two defaults is six months or less is:

$$P(x \leq 6) = 1 - e^{-6/15} = 0.3297$$

Using equation (5.24), we calculate the probability that the time between defaults is 18 months or less:

$$P(x \leq 18) = 1 - e^{-18/15} = 0.6988$$

Thus, the probability that the time between two business loan defaults is between 6 months and 18 months is equal to $0.6988 - 0.3297 = 0.3691$. Probabilities for any other interval can be computed similarly.

Figure 5.31 shows how we can calculate these values for an exponential distribution in Excel using the function EXPON.DIST. The EXPON.DIST function has three inputs: the first input is x, the second input is $1/m$, and the third input is TRUE or FALSE. An input of TRUE for the third input provides the cumulative distribution function value and FALSE provides the probability density function value. Cell B3 calculates $P(x \leq 18)$ using the formula *=EXPON.DIST(18, 1/$B$1, TRUE)*, where cell B1 contains the mean of the exponential distribution. Cell B4 calculates the value for $P(x \leq 6)$ and cell B5 calculates the value for $P(6 \leq x \leq 18) = P(x \leq 18) - P(x \leq 6)$ by subtracting the value in cell B4 from the value in cell B3.

We can calculate P(6 ≤ x ≤ 18) *in a single cell using the formula =EXPON.DIST(18, 1/$B$1, TRUE) - EXPON.DIST(6, 1/$B$1, TRUE).*

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 5.31 Using Excel to Calculate $P(6 \leq x \leq 18)$ for the Time Between Business Loan Defaults Example

	A	B	C
1	Mean, μ =	15	
2			
3	P(x ≤ 18) =	=EXPON.DIST(18, 1/$B$1, TRUE)	
4	P(x ≤ 6) =	=EXPON.DIST(6, 1/$B$1, TRUE)	
5	P(6 ≤ x ≤ 18) = P(x ≤ 18) - P(x ≤ 6) =	=B3 - B4	
6			

	A	B	C
1	Mean, μ =	15	
2			
3	P(x ≤ 18) =	0.6988	
4	P(x ≤ 6) =	0.3297	
5	P(6 ≤ x ≤ 18) = P(x ≤ 18) - P(x ≤ 6) =	0.3691	

Notes + Comments

1. The way we describe probabilities is different for a discrete random variable than it is for a continuous random variable. For discrete random variables, we can talk about the probability of the random variable assuming a particular value. For continuous random variables, we can only talk about the probability of the random variable assuming a value within a given interval.
2. To see more clearly why the height of a probability density function is not a probability, think about a random variable with the following uniform probability distribution:

$$f(x) = \begin{cases} 2 & \text{for } 0 \leq x \leq 0.5 \\ 0 & \text{elsewhere} \end{cases}$$

   The height of the probability density function, $f(x)$, is 2 for values of x between 0 and 0.5. However, we know that probabilities can never be greater than 1. Thus, we see that $f(x)$ cannot be interpreted as the probability of x.
3. The standard normal distribution is the special case of the normal distribution for which the mean is 0 and the standard deviation is 1. This is useful because probabilities for all normal distributions can be computed using the standard normal distribution. We can convert any normal random variable x with mean μ and standard deviation σ to the standard normal random variable z by using the formula $z = \frac{x - \mu}{\sigma}$. We interpret z as the number of standard deviations that the normal random variable x is from its mean μ. Then we can use a table of standard normal probability distributions to find the area under the curve using z and the standard normal probability table. Excel contains special functions for the standard normal distribution: NORM.S.DIST and NORM.S.INV. The function NORM.S.DIST is similar to the function NORM.DIST, but it requires only two input values: the value of interest for calculating the probability and TRUE or FALSE, depending on whether you are interested in finding the probability density or the cumulative distribution function. NORM.S.INV is similar to the NORM.INV function, but it requires only the single input of the probability of interest. Both NORM.S.DIST and NORM.S.INV do not need the additional parameters because they assume a mean of 0 and standard deviation of 1 for the standard normal distribution.
4. A property of the exponential distribution is that the mean and the standard deviation are equal to each other.
5. The (continuous) exponential distribution is related to the (discrete) Poisson distribution. If the Poisson distribution provides an appropriate description of the number of occurrences per interval, the exponential distribution provides a description of the length of the interval between occurrences. This relationship often arises in queueing applications in which, if arrivals follow a Poisson distribution, the time between arrivals must follow an exponential distribution.
6. Chapter 13 explains how values for discrete and continuous random variables can be randomly generated in Excel for use in simulation models.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary

In this chapter we introduced the concept of probability as a means of understanding and measuring uncertainty. Uncertainty is a factor in virtually all business decisions, thus an understanding of probability is essential to modeling such decisions and improving the decision-making process.

We introduced some basic relationships in probability including the concepts of outcomes, events, and calculations of related probabilities. We introduced the concept of conditional probability and discussed how to calculate posterior probabilities from prior probabilities using Bayes' theorem. We then discussed both discrete and continuous random variables as well as some of the more common probability distributions related to these types of random variables. These probability distributions included the custom discrete, discrete uniform, binomial, and Poisson probability distributions for discrete random variables, as well as the uniform, triangular, normal, and exponential probability distributions for continuous random variables. We also discussed the concepts of the expected value (mean) and variance of a random variable.

Probability is used in many chapters that follow in this textbook. In Chapter 6, various measures for showing the strength of association rules are based on probability and conditional probability concepts. Random variables and probability distributions will be seen again in Chapter 7 when we discuss the use of statistical inference to draw conclusions about a population from sample data. In Chapter 8, we will see that the normal distribution is fundamentally involved when we discuss regression analysis as a way of estimating relationships between variables. Chapter 13 demonstrates the use of a variety of probability distributions in simulation models to evaluate the impact of uncertainty on decision-making. Conditional probability and Bayes' theorem will be discussed again in Chapter 17 in the context of decision analysis. It is very important to have a basic understanding of probability, such as is provided in this chapter, as you continue to improve your skills in business analytics.

Glossary

Addition law A probability law used to compute the probability of the union of events. For two events A and B, the addition law is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. For two mutually exclusive events, $P(A \cap B) = 0$, so $P(A \cup B) = P(A) + P(B)$.

Bayes' theorem A method used to compute posterior probabilities.

Binomial probability distribution A probability distribution for a discrete random variable showing the probability of x successes in n trials.

Complement of *A* The event consisting of all outcomes that are not in A.

Conditional probability The probability of an event given that another event has already occurred. The conditional probability of A given B is $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$.

Continuous random variable A random variable that may assume any numerical value in an interval or collection of intervals. An interval can include negative and positive infinity.

Custom discrete probability distribution A probability distribution for a discrete random variable for which each value x_i that the random variable assumes is associated with a defined probability $f(x_i)$.

Discrete random variable A random variable that can take on only specified discrete values.

Discrete uniform probability distribution A probability distribution in which each possible value of the discrete random variable has the same probability.

Empirical probability distribution A probability distribution for which the relative frequency method is used to assign probabilities.

Event A collection of outcomes.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Expected value A measure of the central location, or mean, of a random variable.
Exponential probability distribution A continuous probability distribution that is useful in computing probabilities for the time it takes to complete a task or the time between arrivals. The mean and standard deviation for an exponential probability distribution are equal to each other.
Independent events Two events A and B are independent if $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$; the events do not influence each other.
Intersection of A and B The event containing the outcomes belonging to both A and B. The intersection of A and B is denoted $A \cap B$.
Joint probabilities The probability of two events both occurring; in other words, the probability of the intersection of two events.
Marginal probabilities The values in the margins of a joint probability table that provide the probabilities of each event separately.
Multiplication law A law used to compute the probability of the intersection of events. For two events A and B, the multiplication law is $P(A \cap B) = P(B)P(A \mid B)$ or $P(A \cap B) = P(A)P(B \mid A)$. For two independent events, it reduces to $P(A \cap B) = P(A)P(B)$.
Mutually exclusive events Events that have no outcomes in common; $A \cap B$ is empty and $P(A \cap B) = 0$.
Normal probability distribution A continuous probability distribution in which the probability density function is bell shaped and determined by its mean μ and standard deviation σ.
Poisson probability distribution A probability distribution for a discrete random variable showing the probability of x occurrences of an event over a specified interval of time or space.
Posterior probabilities Revised probabilities of events based on additional information.
Prior probability Initial estimate of the probabilities of events.
Probability A numerical measure of the likelihood that an event will occur.
Probability density function A function used to compute probabilities for a continuous random variable. The area under the graph of a probability density function over an interval represents probability.
Probability distribution A description of how probabilities are distributed over the values of a random variable.
Probability mass function A function, denoted by $f(x)$, that provides the probability that x assumes a particular value for a discrete random variable.
Probability of an event Equal to the sum of the probabilities of outcomes for the event.
Random experiment A process that generates well-defined experimental outcomes. On any single repetition or trial, the outcome that occurs is determined by chance.
Random variables A numerical description of the outcome of an experiment.
Sample space The set of all outcomes.
Standard deviation Positive square root of the variance.
Triangular probability distribution A continuous probability distribution in which the probability density function is shaped like a triangle defined by the minimum possible value a, the maximum possible value b, and the most likely value m. A triangular probability distribution is often used when only subjective estimates are available for the minimum, maximum, and most likely values.
Uniform probability distribution A continuous probability distribution for which the probability that the random variable will assume a value in any interval is the same for each interval of equal length.
Union of A and B The event containing the outcomes belonging to A or B or both. The union of A and B is denoted by $A \cup B$.
Variance A measure of the variability, or dispersion, of a random variable.
Venn diagram A graphical representation of the sample space and operations involving events, in which the sample space is represented by a rectangle and events are represented as circles within the sample space.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Problems

1. **TikTok User Ages.** It is estimated that nearly half of all users of the social networking platform TikTok in the United States were between the ages of 10 and 29 in 2020 (*Statista.com* website). Suppose that the results below were received from a survey sent to a random sample of people in the United States regarding their use of TikTok. **LO 1**

Age Range of Respondent	Uses TikTok	Does Not Use TikTok
10–19	36	44
20–29	87	52
30–39	42	89
40+	44	138
Total:	209	323

   a. Use the sample data above to compute the probability that a respondent to this survey uses TikTok.
   b. What is the probability of a respondent to this survey in each age range (10–19, 20–29, 30–39, 40+) using TikTok? Which age range has the highest probability of a respondent being a user of TikTok?

2. **Airline Performance Measures.** On-time arrivals, lost baggage, and customer complaints are three measures that are typically used to measure the quality of service being offered by airlines. Suppose that the following values represent the on-time arrival percentage, amount of lost baggage, and customer complaints for 10 U.S. airlines. **LO 1, 2**

Airline	On-Time Arrivals (%)	Mishandled Baggage per 1,000 Passengers	Customer Complaints per 1,000 Passengers
Virgin America	83.5	0.87	1.50
JetBlue	79.1	1.88	0.79
AirTran Airways	87.1	1.58	0.91
Delta Air Lines	86.5	2.10	0.73
Alaska Airlines	87.5	2.93	0.51
Frontier Airlines	77.9	2.22	1.05
Southwest Airlines	83.1	3.08	0.25
US Airways	85.9	2.14	1.74
American Airlines	76.9	2.92	1.80
United Airlines	77.4	3.87	4.24

   a. Based on the data above, if you randomly choose a Delta Air Lines flight, what is the probability that this individual flight will have an on-time arrival?
   b. If you randomly choose 1 of the 10 airlines for a follow-up study on airline quality ratings, what is the probability that you will choose an airline with less than two mishandled baggage reports per 1,000 passengers?
   c. If you randomly choose 1 of the 10 airlines for a follow-up study on airline quality ratings, what is the probability that you will choose an airline with more than one customer complaint per 1,000 passengers?
   d. What is the probability that a randomly selected AirTran Airways flight will not arrive on time?

3. **Rolling a Pair of Dice.** Consider the random experiment of rolling a pair of six-sided dice. Suppose that we are interested in the sum of the face values showing on the dice. **LO 1**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. How many outcomes are possible?
b. List the outcomes.
c. What is the probability of obtaining a value of 7?
d. What is the probability of obtaining a value of 9 or greater?

4. **NCAA Basketball Championships**. Only 15 college basketball programs have won more than a single NCAA Basketball Championship. Those 15 colleges are shown below along with their current NCAA Conference Affiliation and the number of NCAA Basketball Championships they have won. Suppose that the winner of one of these 60 NCAA Basketball Championships is selected at random. **LO 1**

DATA*file* basketball

College	Current NCAA Conference Affiliation	Number of NCAA Basketball Championships
UCLA	Pac-12	11
Kentucky	SEC	8
North Carolina	ACC	6
Duke	ACC	5
Indiana	Big Ten	5
Kansas	Big 12	4
UConn	Big East	4
Villanova	Big East	3
Cincinnati	American	2
Florida	SEC	2
Louisville	ACC	2
Michigan State	Big Ten	2
NC State	ACC	2
Oklahoma State	Big 12	2
San Francisco	West Coast	2
		Total: 60

a. What is the probability that the selected winner is from UCLA?
b. What is the probability that the selected winner is from the Big Ten Conference?
c. What is the probability that the selected winner is from the ACC?
d. What is the probability that the selected winners is from either the Big Ten or the ACC?

5. **Ivy League College Admissions.** Suppose that for a recent admissions class, an Ivy League college received 2,851 applications for early admission. Of this group, it admitted 1,033 students early, rejected 854 outright, and deferred 964 to the regular admission pool for further consideration. In the past, this school has admitted 18% of the deferred early admission applicants during the regular admission process. Counting the students admitted early and the students admitted during the regular admission process, the total class size was 2,375. Let E, R, and D represent the events that a student who applies for early admission is admitted early, rejected outright, or deferred to the regular admissions pool. **LO 1, 2**
a. Use the data to estimate $P(E)$, $P(R)$, and $P(D)$.
b. Are events E and D mutually exclusive? Find $P(E \cap D)$.
c. For the 2,375 students who were admitted, what is the probability that a randomly selected student was accepted during early admission?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

d. Suppose a student applies for early admission. What is the probability that the student will be admitted for early admission or be deferred and later admitted during the regular admission process?

6. **Two Events, *A* and *B*.** Suppose that we have two events, A and B, with $P(A) = 0.50$, $P(B) = 0.60$, and $P(A \cap B) = 0.40$. **LO 2, 3, 4**
   a. Find $P(A \mid B)$.
   b. Find $P(B \mid A)$.
   c. Are A and B independent? Why or why not?

7. **Intent to Pursue MBA.** Students taking the Graduate Management Admissions Test (GMAT) were asked about their undergraduate major and intent to pursue their MBA as a full-time or part-time student. A summary of their responses is shown in the table that follows. **LO 3, 4, 5**

		Undergraduate Major			
		Business	Engineering	Other	Totals
Intended Enrollment Status	Full-Time	352	197	251	800
	Part-Time	150	161	194	505
	Totals	502	358	445	1,305

   a. Develop a joint probability table for these data.
   b. Use the marginal probabilities of undergraduate major (business, engineering, or other) to comment on which undergraduate major produces the most potential MBA students.
   c. If a student intends to attend classes full time in pursuit of an MBA degree, what is the probability that the student was an undergraduate engineering major?
   d. If a student was an undergraduate business major, what is the probability that the student intends to attend classes full time in pursuit of an MBA degree?
   e. Let F denote the event that the student intends to attend classes full time in pursuit of an MBA degree, and let B denote the event that the student was an undergraduate business major. Are events F and B independent? Justify your answer.

8. **Student Loans and College Degrees.** More than 40 million Americans are estimated to have at least one outstanding student loan to help pay college expenses (CNNMoney web site). Not all of these graduates pay back their debt in satisfactory fashion. Suppose that the following joint probability table shows the probabilities of student loan status and whether or not the student had received a college degree. **LO 2, 3, 5**

		College Degree		
		Yes	No	
Loan Status	Satisfactory	0.26	0.24	0.50
	Delinquent	0.16	0.34	0.50
		0.42	0.58	

   a. What is the probability that a student with a student loan had received a college degree?
   b. What is the probability that a student with a student loan had not received a college degree?
   c. Given that the student has received a college degree, what is the probability that the student has a delinquent loan?
   d. Given that the student has not received a college degree, what is the probability that the student has a delinquent loan?
   e. What is the impact of dropping out of college without a degree for students who have a student loan?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
optilytics

9. **Senior Data Scientist Position Applicants.** The Human Resources Manager for Optilytics LLC is evaluating applications for the position of Senior Data Scientist. The file *optilytics* presents summary data of the applicants for the position. The collected data include the highest degree achieved and the years of work experience for each candidate as well as whether or not the candidate has experience using Alteryx, a software used for analytics. **LO 2, 3, 5**
   a. Use a PivotTable in Excel to create a joint probability table showing the probabilities associated with a randomly selected applicant's use of Alteryx and highest degree achieved. Use this joint probability table to answer the questions below.
   b. What are the marginal probabilities? What do they tell you about the probabilities associated with the use of Alteryx among the applicants and highest degree completed by applicants?
   c. If the applicant has not used Alteryx, what is the probability that the highest degree completed by the applicant is a PhD?
   d. If the highest degree completed by the applicant is a bachelor's degree, what is the probability that the applicant has used Alteryx?
   e. What is the probability that a randomly selected applicant has used Alteryx and whose highest completed degree is a PhD?

10. **U.S. Household Incomes.** The U.S. Census Bureau is a leading source of quantitative data related to the people and economy of the United States. The crosstabulation below represents the number of households (thousands) and the household income by the highest level of education for the head of household (U.S. Census Bureau web site). Use this crosstabulation to answer the following questions. **LO 2, 3, 5**

	Household Income				
Highest Level of Education	Under $25,000	$25,000–$49,999	$50,000–$99,999	$100,000 and Over	Total
High school graduate	9,880	9,970	9,441	3,482	32,773
Bachelor's degree	2,484	4,164	7,666	7,817	22,131
Master's degree	685	1,205	3,019	4,094	9,003
Doctoral degree	79	160	422	1,076	1,737
Total	13,128	15,499	20,548	16,469	65,644

   a. Develop a joint probability table.
   b. What is the probability the head of one of these households has a master's degree or higher education?
   c. What is the probability a household is headed by someone with a high school diploma earning $100,000 or more?
   d. What is the probability one of these households has an income below $25,000?
   e. What is the probability a household is headed by someone with a bachelor's degree earning less than $25,000?
   f. Are household income and educational level independent?

11. **Probability of Homes Selling.** Cooper Realty is a small real estate company located in Albany, New York, that specializes primarily in residential listings. The company recently became interested in determining the likelihood of one of its listings being sold within a certain number of days. An analysis of company sales of 800 homes in previous years produced the following data. **LO 1, 2, 3, 4, 5**
   a. If A is defined as the event that a home is listed for more than 90 days before being sold, estimate the probability of A.
   b. If B is defined as the event that the initial asking price is under $150,000, estimate the probability of B.
   c. What is the probability of $A \cap B$?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

d. Assuming that a contract was just signed to list a home with an initial asking price of less than $150,000, what is the probability that the home will take Cooper Realty more than 90 days to sell?
e. Are events A and B independent?

		Days Listed Until Sold			
		Under 30	31–90	Over 90	Total
Initial Asking Price	Under $150,000	50	40	10	100
	$150,000–$199,999	20	150	80	250
	$200,000–$250,000	20	280	100	400
	Over $250,000	10	30	10	50
	Total	100	500	200	800

12. **Computing Probabilities.** The prior probabilities for events A_1 and A_2 are $P(A_1) = 0.40$ and $P(A_2) = 0.60$. It is also known that $P(A_1 \cap A_2) = 0$. Suppose $P(B \mid A_1) = 0.20$ and $P(B \mid A_2) = 0.05$. **LO 2, 3, 6**
    a. Are A_1 and A_2 mutually exclusive? Explain.
    b. Compute $P(A_1 \cap B)$ and $P(A_2 \cap B)$.
    c. Compute $P(B)$.
    d. Apply Bayes' theorem to compute $P(A_1 \mid B)$ and $P(A_2 \mid B)$.
13. **COVID-19 Testing**. More than 1.5 million tests for the COVID virus were performed each day in the United States during the COVID-19 pandemic in 2021. Many employers, schools, and government agencies required negative COVID-19 tests before allowing people on their premises. Suppose that for a particular COVID-19 test, the probability of a false negative (meaning that a person who is actually infected with the COVID-19 virus tests negative for the virus) is 0.1, and that the test always gives a negative result for someone who is not infected with COVID-19. Suppose also the probability that a randomly selected person is infected with COVID-19 is 0.05. Answer the following questions. **LO 2, 3, 6**
    a. What is the probability that someone who tests negative for the COVID-19 virus actually has COVID-19?
    b. Suppose that a large school system states that 2,500 of its students have reported a negative COVID-19 test. Of these 2,500 students who tested negative, how many would you expect to have COVID-19?
    c. Suppose the incidence of COVID-19 is actually considerably higher and the probability of any randomly selected person having COVID-19 is 0.20. What is the probability in this case that someone who tests negative for the COVID-19 virus actually has COVID-19? Explain this change from part (a).
14. **Credit Card Defaults.** A local bank reviewed its credit-card policy with the intention of recalling some of its credit cards. In the past, approximately 5% of cardholders defaulted, leaving the bank unable to collect the outstanding balance. Hence, management established a prior probability of 0.05 that any particular cardholder will default. The bank also found that the probability of missing a monthly payment is 0.20 for customers who do not default. Of course, the probability of missing a monthly payment for those who default is 1. **LO 2, 3, 6**
    a. Given that a customer missed a monthly payment, compute the posterior probability that the customer will default.
    b. The bank would like to recall its credit card if the probability that a customer will default is greater than 0.20. Should the bank recall its credit card if the customer misses a monthly payment? Why or why not?
15. **Prostate Cancer Screening.** According to an article in *Esquire* magazine, approximately 70% of males over age 70 will develop cancerous cells in their prostate. Prostate cancer is second only to skin cancer as the most common form of cancer for males

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

in the United States. One of the most common tests for the detection of prostate cancer is the prostate-specific antigen (PSA) test. However, this test is known to have a high false-positive rate (tests that come back positive for cancer when no cancer is present). Suppose there is a 0.02 probability that a male patient has prostate cancer before testing. The probability of a false-positive test is 0.75, and the probability of a false-negative (no indication of cancer when cancer is actually present) is 0.20. **LO 2, 3, 6**

a. What is the probability that the male patient has prostate cancer if the PSA test comes back positive?
b. What is the probability that the male patient has prostate cancer if the PSA test comes back negative?
c. For older men, the prior probability of having cancer increases. Suppose that the prior probability of the male patient is 0.3 rather than 0.02. What is the probability that the male patient has prostate cancer if the PSA test comes back positive? What is the probability that the male patient has prostate cancer if the PSA test comes back negative?
d. What can you infer about the PSA test from the results of parts (a), (b), and (c)?

16. **Finding Oil in Alaska.** An oil company purchased an option on land in Alaska. Preliminary geologic studies assigned the following prior probabilities. **LO 2, 3, 6**

$$P(\text{high-quality oil}) = 0.50$$
$$P(\text{medium-quality oil}) = 0.20$$
$$P(\text{no oil}) = 0.30$$

a. What is the probability of finding oil?
b. After 200 feet of drilling on the first well, a soil test is taken. The probabilities of finding the particular type of soil identified by the test are as follows.

$$P(\text{soil} \mid \text{high-quality oil}) = 0.20$$
$$P(\text{soil} \mid \text{medium-quality oil}) = 0.80$$
$$P(\text{soil} \mid \text{no oil}) = 0.20$$

How should the firm interpret the soil test? What are the revised probabilities, and what is the new probability of finding oil?

17. **Unemployment Data.** Suppose the following data represent the number of persons unemployed for a given number of months in Killeen, Texas. The values in the first column show the number of months unemployed and the values in the second column show the corresponding number of unemployed persons. **LO 7**

Months Unemployed	Number Unemployed
1	1,029
2	1,686
3	2,269
4	2,675
5	3,487
6	4,652
7	4,145
8	3,587
9	2,325
10	1,120

Let x be a random variable indicating the number of months a randomly selected person is unemployed.

a. Use the data to develop an empirical discrete probability distribution for x.
b. Show that your probability distribution satisfies the conditions for a valid discrete probability distribution.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. What is the probability that a person is unemployed for two months or less? Unemployed for more than two months?
d. What is the probability that a person is unemployed for more than six months?

18. **Information Systems Job Satisfaction.** The percent frequency distributions of job satisfaction scores for a sample of information systems (IS) senior executives and middle managers are as follows. The scores range from a low of 1 (very dissatisfied) to a high of 5 (very satisfied). **LO 7**

Job Satisfaction Score	IS Senior Executives (%)	IS Middle Managers (%)
1	5	4
2	9	10
3	3	12
4	42	46
5	41	28

a. Develop a probability distribution for the job satisfaction score of a randomly selected senior executive.
b. Develop a probability distribution for the job satisfaction score of a randomly selected middle manager.
c. What is the probability that a randomly selected senior executive will report a job satisfaction score of 4 or 5?
d. What is the probability that a randomly selected middle manager is very satisfied?
e. Compare the overall job satisfaction of senior executives and middle managers.

19. **Expectation and Variance of a Random Variable.** The following table provides a probability distribution for the random variable y. **LO 8**
a. Compute $E(y)$.
b. Compute Var(y) and σ.

y	$f(y)$
2	0.20
4	0.30
7	0.40
8	0.10

DATA *file*
internalaudit

20. **Financial Statement Audits**. Internal auditors are often used to review an organization's financial statements such as balance sheets, income statements, and cash flow statements prior to public filings. Auditors seek to verify that the financial statements accurately represent the financial position of the organization and that the statements follow accepted accounting principles. Many errors that are discovered by auditors are minor errors that are easily corrected. However, some errors are serious and require substantial time to rectify. Suppose that the financial statements of 567 public companies are audited. The file *internalaudit* contains the number of errors discovered during the internal audit of each of these 567 public companies that were classified as "serious" errors. Use the data in the file *internalaudit* to answer the following. **LO 7, 8**
a. Construct an empirical discrete probability distribution for the number of serious errors discovered during the internal audits of these 567 public companies.
b. What is the probability that a company has no serious errors in its financial statements?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. What is the probability that a company has four or more serious errors in its financial statements?
d. What is the expected number of serious errors in a company's financial statements?
e. What is the variance of the number of serious errors in a company's financial statements?
f. What is the standard deviation of the number of serious errors in a company's financial statements?

21. **Damage Claims at an Insurance Company.** The probability distribution for damage claims paid by the Newton Automobile Insurance Company on collision insurance is as follows. LO 8

Payment ($)	Probability
0	0.85
500	0.04
1,000	0.04
3,000	0.03
5,000	0.02
8,000	0.01
10,000	0.01

a. Use the expected collision payment to determine the collision insurance premium that would enable the company to break even.
b. The insurance company charges an annual rate of $520 for the collision coverage. What is the expected value of the collision policy for a policyholder? (*Hint:* It is the expected payments from the company minus the cost of coverage.) Why does the policyholder purchase a collision policy with this expected value?

22. **Plant Expansion Decision.** The J.R. Ryland Computer Company is considering a plant expansion to enable the company to begin production of a new computer product. The company's president must determine whether to make the expansion a medium- or large-scale project. Demand for the new product is uncertain, which for planning purposes may be low demand, medium demand, or high demand. The probability estimates for demand are 0.20, 0.50, and 0.30, respectively. Letting x and y indicate the annual profit in thousands of dollars, the firm's planners developed the following profit forecasts for the medium- and large-scale expansion projects. LO 8

		Medium-Scale Expansion Profit		Large-Scale Expansion Profit	
		x	$f(x)$	y	$f(y)$
Demand	Low	50	0.20	0	0.20
	Medium	150	0.50	100	0.50
	High	200	0.30	300	0.30

a. Compute the expected value for the profit associated with the two expansion alternatives. Which decision is preferred for the objective of maximizing the expected profit?
b. Compute the variance for the profit associated with the two expansion alternatives. Which decision is preferred for the objective of minimizing the risk or uncertainty?

23. **Binomial Distribution Calculations.** Consider a binomial experiment with $n = 10$ and $p = 0.10$. LO 9
a. Compute $f(0)$.
b. Compute $f(2)$.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Compute $P(x \leq 2)$.
d. Compute $P(x \geq 1)$.
e. Compute $E(x)$.
f. Compute Var(x) and σ.

24. **Acceptance Sampling.** Many companies use a quality control technique called acceptance sampling to monitor incoming shipments of parts, raw materials, and so on. In the electronics industry, component parts are commonly shipped from suppliers in large lots. Inspection of a sample of n components can be viewed as the n trials of a binomial experiment. The outcome for each component tested (trial) will be that the component is classified as good or defective. Reynolds Electronics accepts a lot from a particular supplier if the defective components in the lot do not exceed 1%. Suppose a random sample of five items from a recent shipment is tested. **LO 9**
    a. Assume that 1% of the shipment is defective. Compute the probability that no items in the sample are defective.
    b. Assume that 1% of the shipment is defective. Compute the probability that exactly one item in the sample is defective.
    c. What is the probability of observing one or more defective items in the sample if 1% of the shipment is defective?
    d. Would you feel comfortable accepting the shipment if one item was found to be defective? Why or why not?

25. **Dog Food Marketing Focus Group.** According to the American Veterinary Medical Association (AVMA), 38.4% of households in the United States own a dog as a pet (AVMA website). Suppose that a company that sells dog food would like to establish a focus group to gather input on a new dog food marketing campaign. The company plans to contact 25 randomly selected households to invite people to join the focus group. **LO 9**
    a. Compute the probability that 10 of these 25 households own a dog as a pet.
    b. Compute the probability that 2 or fewer of these 25 households own a dog as a pet.
    c. For the sample of 25 households, compute the expected number of households who own a dog as a pet.
    d. For the sample of 25 households, compute the variance and standard deviation of households who own a dog as a pet.

26. **Introductory Statistics Course Withdrawals.** A university found that 20% of its students withdraw without completing the introductory statistics course. Assume that 20 students registered for the course. **LO 9**
    a. Compute the probability that two or fewer students will withdraw.
    b. Compute the probability that exactly four students will withdraw.
    c. Compute the probability that more than three students will withdraw.
    d. Compute the expected number of student withdrawals.

27. **Poisson Distribution Calculations.** Consider a Poisson distribution with $\mu = 3$.
    a. Write the appropriate Poisson probability mass function. **LO 10**
    b. Compute $f(2)$.
    c. Compute $f(1)$.
    d. Compute $P(x \geq 2)$.

28. **Computer Code Errors.** The book *Code Complete* by Steve McDonnell estimates that there are 15 to 50 errors per 1,000 lines of delivered code for computer programs. Assume that for a particular software package, the error rate is 25 per 1,000 lines of code and that the number of errors per 1,000 lines of code follows a Poisson distribution. **LO 10**
    a. What is the probability that a portion of code for this software package that contains 250 lines of code contains no errors?
    b. What is the probability of there being exactly 30 errors in 1,000 lines of code?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. What is the probability of more than 3 errors in 100 lines of code?
d. What is the expected number of errors in a portion of code that contains 25,000 lines of code?

29. **911 Calls.** Emergency 911 calls in a large metropolitan area come in at the rate of one every 2 minutes. Assume that the number of 911 calls is a random variable that can be described by the Poisson distribution. **LO 10**
a. What is the expected number of 911 calls in 1 hour?
b. What is the probability of three 911 calls in 5 minutes?
c. What is the probability of no 911 calls during a 5-minute period?

30. **Small Business Failures.** A regional director responsible for business development in the state of Pennsylvania is concerned about the number of small business failures. If the mean number of small business failures per month is 10 and the number of failures follows a Poisson distribution, what is the probability that exactly 4 small businesses will fail during a given month? **LO 10**

31. **Uniform Distribution Calculations.** The random variable x is known to be uniformly distributed between 10 and 20. **LO 11**
a. Show the graph of the probability density function.
b. Compute $P(x < 15)$.
c. Compute $P(12 \leq x \leq 18)$.
d. Compute $E(x)$.
e. Compute Var(x).

32. **RAND Function in Excel.** Most computer languages include a function that can be used to generate random numbers. In Excel, the RAND function can be used to generate random numbers between 0 and 1. If we let x denote a random number generated using RAND, then x is a continuous random variable with the probability density function that follows. **LO 11**

$$f(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

a. Graph the probability density function.
b. What is the probability of generating a random number between 0.25 and 0.75?
c. What is the probability of generating a random number with a value less than or equal to 0.30?
d. What is the probability of generating a random number with a value greater than 0.60?
e. Generate 50 random numbers by entering =RAND() into 50 cells of an Excel worksheet.
f. Compute the mean and standard deviation for the random numbers in part (e).

33. **Bidding on a Piece of Land.** Suppose we are interested in bidding on a piece of land and we know one other bidder is interested. The seller announced that the highest bid in excess of $10,000 will be accepted. Assume that the competitor's bid x is a random variable that is uniformly distributed between $10,000 and $15,000. **LO 11**
a. Suppose you bid $12,000. What is the probability that your bid will be accepted?
b. Suppose you bid $14,000. What is the probability that your bid will be accepted?
c. What amount should you bid to maximize the probability that you get the property?
d. Suppose you know someone who is willing to pay you $16,000 for the property. Would you consider bidding less than the amount in part (c)? Why or why not?

34. **Triangular Distribution Calculations.** A random variable has a triangular probability density function with $a = 50$, $b = 375$, and $m = 250$. **LO 12**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Sketch the probability distribution function for this random variable. Label the points $a = 50$, $b = 375$, and $m = 250$ on the x-axis.
b. What is the probability that the random variable will assume a value between 50 and 250?
c. What is the probability that the random variable will assume a value greater than 300?

35. **Project Completion Time.** The Siler Construction Company is about to bid on a new industrial construction project. To formulate their bid, the company needs to estimate the time required for the project. Based on past experience, management expects that the project will require at least 24 months, and could take as long as 48 months if there are complications. The most likely scenario is that the project will require 30 months. Assume that the actual time for the project can be approximated using a triangular probability distribution. **LO 12**
    a. What is the probability that the project will take less than 30 months?
    b. What is the probability that the project will take between 28 and 32 months?
    c. To submit a competitive bid, the company believes that if the project takes more than 36 months, then the company will lose money on the project. Management does not want to bid on the project if there is greater than a 25% chance that they will lose money on this project. Should the company bid on this project?
36. **Large-Cap Stock Fund Returns.** Suppose that the return for a particular large-cap stock fund is normally distributed with a mean of 14.4% and standard deviation of 4.4%. **LO 13**
    a. What is the probability that the large-cap stock fund has a return of at least 20%?
    b. What is the probability that the large-cap stock fund has a return of 10% or less?
37. **IQ Scores and Mensa.** A person must score in the upper 2% of the population on an IQ test to qualify for membership in Mensa, the international high IQ society. If IQ scores are normally distributed with a mean of 100 and a standard deviation of 15, what score must a person have to qualify for Mensa? **LO 14**
38. **Amount of Sleep**. The United States Centers for Disease Control and Prevention (CDC) recommends that adults sleep seven to nine hours per night (*CDC.gov* website). However, many adults in the United States sleep less than seven hours per night. Suppose that the amount of sleep per night for an adult in the United States follows a normal distribution with a mean of 8.0 hours and standard deviation of 1.7 hours. **LO 13, 14**
    a. What is the probability that a randomly selected adult in the United States sleeps less than 7.0 hours per night?
    b. What is the probability that a randomly selected adult in the United States sleeps between 7.0 and 9.0 hours per night?
    c. How many hours per night would an adult in the United States sleep if they are in the 10th percentile of this distribution?
39. **Web Site Traffic.** Assume that the traffic to the web site of Smiley's People, Inc., which sells customized T-shirts, follows a normal distribution, with a mean of 4.5 million visitors per day and a standard deviation of 820,000 visitors per day. **LO 13, 14**
    a. What is the probability that the web site has fewer than 5 million visitors in a single day?
    b. What is the probability that the web site has 3 million or more visitors in a single day?
    c. What is the probability that the web site has between 3 million and 4 million visitors in a single day?
    d. Assume that 85% of the time, the Smiley's People web servers can handle the daily web traffic volume without purchasing additional server capacity. What is the amount of web traffic that will require Smiley's People to purchase additional server capacity?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

40. **Probability of Defect.** Suppose that Motorola uses the normal distribution to determine the probability of defects and the number of defects in a particular production process. Assume that the production process manufactures items with a mean weight of 10 ounces. **LO 13, 14**
    a. Assume that the process standard deviation is 0.15, and the process control is set at plus or minus one standard deviation. Units with weights less than 9.85 or greater than 10.15 ounces will be classified as defects. Calculate the probability of a defect and the expected number of defects for a 1,000-unit production run.
    b. Assume that through process design improvements, the process standard deviation can be reduced to 0.05. Assume that the process control remains the same, with weights less than 9.85 or greater than 10.15 ounces being classified as defects. Calculate the probability of a defect and the expected number of defects for a 1,000-unit production run.
    c. What is the advantage of reducing process variation, thereby causing process control limits to be at a greater number of standard deviations from the mean?

41. **Exponential Distribution Calculations.** Consider the exponential probability density function that follows. **LO 15**

$$f(x) = \frac{1}{3}e^{-x/3} \quad \text{for } x \geq 0$$

    a. Write the formula for $P(x \leq x_0)$.
    b. Find $P(x \leq 2)$.
    c. Find $P(x \geq 3)$.
    d. Find $P(x \leq 5)$.
    e. Find $P(2 \leq x \leq 5)$.

42. **Vehicle Arrivals at an Intersection.** The time between arrivals of vehicles at a particular intersection follows an exponential probability distribution with a mean of 12 seconds. **LO 15**
    a. Sketch this exponential probability distribution.
    b. What is the probability that the arrival time between vehicles is 12 seconds or less?
    c. What is the probability that the arrival time between vehicles is 6 seconds or less?
    d. What is the probability of 30 or more seconds between vehicle arrivals?

43. **Time Spent Playing *World of Warcraft*.** Suppose that the time spent by players in a single session on the *World of Warcraft* multiplayer online role-playing game follows an exponential distribution with a mean of 38.3 minutes. **LO 15**
    a. Write the exponential probability distribution function for the time spent by players on a single session of *World of Warcraft*.
    b. What is the probability that a player will spend between 20 and 40 minutes on a single session of *World of Warcraft*?
    c. What is the probability that a player will spend more than 1 hour on a single session of *World of Warcraft*?

44. **Laffy Taffy Candy Production.** Laffy Taffy is a type of taffy candy made from corn syrup, sugar, palm oil, and other ingredients. Laffy Taffy comes in a variety of flavors including strawberry, banana, and cherry. Suppose that Laffy Taffy is produced as a continuous length of taffy on an extrusion machine. Defects can occur in the taffy including large air bubbles and tears. Assume that the distance between such defects follows an exponential distribution with a mean of 523 meters. **LO 15**
    a. What is the probability that the machine will produce more than 1,000 meters of taffy before the next defect is observed?
    b. What is the probability that the next defect will be observed before the machine produces 500 meters or less of taffy?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Case Problem 1: Hamilton County Judges

DATA*file* judge

Hamilton County judges try thousands of cases per year. In an overwhelming majority of the cases disposed, the verdict stands as rendered. However, some cases are appealed, and of those appealed, some of the cases are reversed. Kristen DelGuzzi of the *Cincinnati Enquirer* newspaper conducted a study of cases handled by Hamilton County judges over a three-year period. Shown in the table below, and in the file *judge*, are the results for 182,908 cases handled (disposed) by 38 judges in Common Pleas Court, Domestic Relations Court, and Municipal Court. Two of the judges (Dinkelacker and Hogan) did not serve in the same court for the entire three-year period.

The purpose of the newspaper's study was to evaluate the performance of the judges. Appeals are often the result of mistakes made by judges, and the newspaper wanted to know which judges were doing a good job and which were making too many mistakes. You are called in to assist in the data analysis. Use your knowledge of probability and conditional probability to help with the ranking of the judges. You also may be able to analyze the likelihood of appeal and reversal for cases handled by different courts. **LO 1, 3**

Total Cases Disposed, Appealed, and Reversed in Hamilton County Courts

Common Pleas Court

Judge	Total Cases Disposed	Appealed Cases	Reversed Cases
Fred Cartolano	3,037	137	12
Thomas Crush	3,372	119	10
Patrick Dinkelacker	1,258	44	8
Timothy Hogan	1,954	60	7
Robert Kraft	3,138	127	7
William Mathews	2,264	91	18
William Morrissey	3,032	121	22
Norbert Nadel	2,959	131	20
Arthur Ney, Jr.	3,219	125	14
Richard Niehaus	3,353	137	16
Thomas Nurre	3,000	121	6
John O'Connor	2,969	129	12
Robert Ruehlman	3,205	145	18
J. Howard Sundermann	955	60	10
Ann Marie Tracey	3,141	127	13
Ralph Winkler	3,089	88	6
Total	43,945	1,762	199

Domestic Relations Court

Judge	Total Cases Disposed	Appealed Cases	Reversed Cases
Penelope Cunningham	2,729	7	1
Patrick Dinkelacker	6,001	19	4
Deborah Gaines	8,799	48	9
Ronald Panioto	12,970	32	3
Total	30,499	106	17

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Municipal Court

Judge	Total Cases Disposed	Appealed Cases	Reversed Cases
Mike Allen	6,149	43	4
Nadine Allen	7,812	34	6
Timothy Black	7,954	41	6
David Davis	7,736	43	5
Leslie Isaiah Gaines	5,282	35	13
Karla Grady	5,253	6	0
Deidra Hair	2,532	5	0
Dennis Helmick	7,900	29	5
Timothy Hogan	2,308	13	2
James Patrick Kenney	2,798	6	1
Joseph Luebbers	4,698	25	8
William Mallory	8,277	38	9
Melba Marsh	8,219	34	7
Beth Mattingly	2,971	13	1
Albert Mestemaker	4,975	28	9
Mark Painter	2,239	7	3
Jack Rosen	7,790	41	13
Mark Schweikert	5,403	33	6
David Stockdale	5,371	22	4
John A. West	2,797	4	2
Total	108,464	500	104

Managerial Report

Prepare a report with your rankings of the judges. Also, include an analysis of the likelihood of appeal and case reversal in the three courts. At a minimum, your report should include the following:

1. The probability of cases being appealed and reversed in the three different courts.
2. The probability of a case being appealed for each judge.
3. The probability of a case being reversed for each judge.
4. The probability of reversal given an appeal for each judge.
5. Rank the judges within each court. State the criteria you used and provide a rationale for your choice.

Case Problem 2: McNeil's Auto Mall

Harriet McNeil, proprietor of McNeil's Auto Mall, believes that it is good business for the automobile dealership to have more customers on the lot than can be served, as this creates an impression that demand for the automobiles on the lot is high. However, Harriet also understands that if there are far more customers on the lot than can be served by sales people, the dealership may lose sales to customers who become frustrated and leave without making a purchase.

Harriet is primarily concerned about the staffing of salespeople on the lot on Saturday mornings (8:00 A.M. to noon), which are the busiest time of the week for McNeil's Auto Mall. On Saturday mornings, an average of 6.8 customers arrive per hour. The customers arrive randomly at a constant rate throughout the morning, and a salesperson spends an average of one hour with a customer. Harriet's experience has led to the conclusion that if there are two more customers on the lot than can be served at any time on a Saturday morning, McNeil's Auto Mall achieves the optimal balance of creating an impression of

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

high demand without losing too many customers who become frustrated and leave without making a purchase.

Harriet now wants to determine how many salespeople to have on the lot on Saturday mornings in order to achieve the goal of having two more customers on the lot than can be served at any time. Harriet understands that occasionally the number of customers on the lot will exceed the number of salespersons by more than two, and Harriet is willing to accept such an occurrence no more than 10% of the time. **LO 10**

Managerial Report

Harriet has asked you to determine the number of salespersons to have on the lot on Saturday mornings in order to satisfy the stated criteria. In answering Harriet's question, consider the following three questions:

1. How is the number of customers who arrive to the lot on a Saturday morning distributed?
2. Suppose Harriet currently uses five salespeople on the lot on Saturday mornings. Using the probability distribution you identified in (1), what is the probability that the number of customers who arrive to McNeil's Auto Mall will exceed the number of salespersons by more than two? Does Harriet's current Saturday morning employment strategy satisfy the stated objective? Why or why not?
3. What is the minimum number of salespeople Harriet should have on the lot on Saturday mornings to achieve the stated objective?

Case Problem 3: Gebhardt Electronics

Gebhardt Electronics produces a wide variety of transformers that it sells directly to manufacturers of electronics equipment. For one component used in several models of its transformers, Gebhardt uses a 1-meter length of 0.20 mm diameter solid wire made of pure Oxygen-Free Electronic (OFE) copper. A flaw in the wire reduces its conductivity and increases the likelihood it will break, and this critical component is difficult to reach and repair after a transformer has been constructed. Therefore, Gebhardt wants to use primarily flawless lengths of wire in making this component. The company is willing to accept no more than a 1 in 20 chance that a 1-meter length taken from a spool will be flawless. Gebhardt also occasionally uses smaller pieces of the same wire in the manufacture of other components, so the 1-meter segments to be used for this component are essentially taken randomly from a long spool of 0.20 mm diameter solid OFE copper wire.

Gebhardt is now considering a new supplier for copper wire. This supplier claims that its spools of 0.20 mm diameter solid OFE copper wire average 127 centimeters between flaws. Gebhardt now must determine whether the new supply will be satisfactory if the supplier's claim is valid. **LO 15**

Managerial Report

In making this assessment for Gebhardt Electronics, consider the following three questions:

1. If the new supplier does provide spools of 0.20 mm solid OFE copper wire that average 127 centimeters between flaws, how is the length of wire between two consecutive flaws distributed?
2. Using the probability distribution you identified in (1), what is the probability that Gebhardt's criteria will be met (i.e., a 1 in 20 chance that a randomly selected 1-meter segment of wire provided by the new supplier will be flawless)?
3. In centimeters, what is the minimum mean length between consecutive flaws that would result in satisfaction of Gebhardt's criteria?
4. In centimeters, what is the minimum mean length between consecutive flaws that would result in a 1 in 100 chance that a randomly selected 1-meter segment of wire provided by the new supplier will be flawless?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 6

Descriptive Data Mining

Contents

Analytics in Action:
Advice from a Machine

6.1 Dimension Reduction
- Geometric Interpretation of Principal Component Analysis
- Summarizing Protein Consumption for Maillard Riposte

6.2 Cluster Analysis
- Measuring Distance Between Observations Consisting of Quantitative Variables
- Measuring Distance Between Observations Consisting of Categorical Variables
- *k*-Means Clustering
- Hierarchical Clustering and Measuring Dissimilarity Between Clusters
- Hierarchical Clustering versus *k*-Means Clustering

6.3 Association Rules
- Evaluating Association Rules

6.4 Text Mining
- Voice of the Customer at Triad Airlines
- Preprocessing Text Data for Analysis
- Movie Reviews
- Computing Dissimilarity Between Documents
- Word Clouds

Summary 295
Glossary 296
Problems 298

Available in the Cengage eBook:
- Appendix: Principal Component Analysis with R
- Appendix: *k*-Means Clustering with R
- Appendix: Hierarchical Clustering with R
- Appendix: Association Rules with R
- Appendix: Text Mining with R
- Appendix: Principal Component Analysis with Orange
- Appendix: *k*-Means Clustering with Orange
- Appendix: Hierarchical Clustering with Orange
- Appendix: Association Rules with Orange
- Appendix: Text Mining with Orange

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Employ principal component analysis and interpret principal components.

LO 2 Describe and apply distance measures for observations consisting of quantitative variables.

LO 3 Describe and apply distance measures for observations consisting of binary variables.

LO 4 Construct clusters with *k*-means clustering and assess their viability.

LO 5 Describe and apply agglomerative linkage methods for combining groups of observations.

LO 6 Construct clusters with agglomerative hierarchical clustering and assess their viability.

LO 7 Construct and evaluate association rules from transactional data.

LO 8 Describe and perform text mining using a bag-of-words approach.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

Advice from a Machine

The proliferation of data and increase in computing power have sparked the development of automated *recommender systems*, which provide consumers with suggestions for movies, music, books, clothes, restaurants, dating, and whom to follow on social media platforms. The sophisticated, proprietary algorithms guiding recommender systems measure the degree of similarity between users or items to identify recommendations of potential interest to a user.

Netflix, a company that provides subscription streaming for video content in addition to being a production company, provides its users with recommendations for video content based on each user's expressed interests and feedback on previously viewed content. Netflix is able to gather substantial data on its customers' viewing habits which allows Netflix's recommendations to account for differences in viewing behavior based on the day of the week, the time of day, the device used (computer, phone, television), and even the viewing location.

The use of recommender systems is prevalent in e-commerce and in social media. Using attributes detailed by the Music Genome Project, Pandora Internet Radio plays songs with properties similar to songs that a user "likes." In the online dating world, web sites such as eHarmony, Match.com, and OKCupid use different "formulas" to take into account hundreds of different behavioral traits to propose date "matches." Stitch Fix, a personal shopping service, combines recommendation algorithms and human input from its fashion experts to match its inventory of fashion items to its clients. Facebook uses recommender systems to suggest "friends" to connect with and to populate the user's feed with news stories. YouTube uses recommender systems to suggest videos to the user, and LinkedIn suggests jobs that may be of interest to a user based on the user's characteristics and past search history on the site.

Recommender systems are not without problems. Recommender systems can provide biased results if the data used to "train" the system is incomplete or not representative of the full population of users. Recommender systems can also reinforce existing prejudices and fail to present multiple options to the user. These are all challenging issues that analytics professionals must confront constantly.

Over the past few decades, technological advances have led to a dramatic increase in the amount of recorded data. The use of smartphones, global positioning satellite (GPS) devices, radio-frequency identification (RFID) tags, electronic sensors, credit cards, and the Internet has facilitated the collection of data from phone conversations, e-mails, social media posts, business transactions, product and customer tracking, business transactions, and web browsing. The increase in the use of data-mining techniques in business has been caused largely by three events: the explosion in the amount of data being produced and electronically tracked, the ability to electronically warehouse these data, and the affordability of computer power to analyze the data. In this chapter, we discuss the analysis of large quantities of data in order to gain insight on customers and to uncover patterns to improve business processes.

We define an **observation**, or **record**, as the set of recorded values of variables associated with a single entity. An observation is often displayed as a row of values in a spreadsheet or database in which the columns correspond to the variables. For example, in a university's database of alumni, an observation may correspond to an alumnus's age, marital status, employer, position title, as well as size and frequency of donations to the university.

In this chapter, we focus on descriptive data-mining methods, also called **unsupervised learning** techniques. In an unsupervised learning application, there is no outcome variable to predict; rather, the goal is to use the variable values to identify relationships between observations. Because unsupervised learning approaches are capable of describing patterns and relationships in large data sets with many observations of several variables, these approaches can be thought of as high-dimensional descriptive analytics. Without an explicit outcome (or one that is objectively known), there is no definitive measure of accuracy. Instead, qualitative assessments, such as the interpretability of the results or how well they match expert judgment, are often used to assess the results from an unsupervised learning method.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6.1 Dimension Reduction

Data sets with a large number of variables can be difficult to analyze. Even exploratory data analysis of extremely "wide" data with many variables can be difficult. Data visualizations are typically restricted to the consideration of two or three variables at a time, and if there are dozens of variables, the number of variable combinations to visualize can be overwhelming. Further, both descriptive and predictive data mining algorithms typically work better if the number of variables under consideration is not too large.

We discuss data visualizations in Chapter 3.

Dimension reduction refers to the set of mathematical techniques that reduce the number of variables under consideration by replacing them with new "meta-variables" that are weighted combinations of the original variables. The goal of dimension reduction is to describe a data set with fewer appropriately constructed meta-variables that still allow the analyst to capture the patterns and insight contained in the original variables.

We discuss feature (variable) selection in more detail within the context of linear regression in Chapter 8, and in the context of predictive data mining methods in Chapters 10 and 11.

Dimension reduction techniques differ from **feature (variable) selection**, which is the process of reducing the number of variables under consideration by selecting a subset of the original variables. To identify redundant and irrelevant variables, feature selection often relies on descriptive statistics related to the goals of the data-mining task and the assessment of a specific data mining algorithm's performance when considering various combinations of variables. In contrast, dimension reduction does not need to consider the ultimate data-mining task in order to transform a higher-dimensional data set into a lower-dimensional data set.

In this section, we present **principal component analysis**, a statistical procedure commonly used to execute dimension reduction. Principal component analysis is appropriate for reducing the dimension of data sets consisting of quantitative variables. We begin with a low-dimensional example to provide a geometric interpretation of principal component analysis. Then, we demonstrate the application of principle component analysis on a higher-dimensional example.

Geometric Interpretation of Principal Component Analysis

DATA*file*
states.csv

Suppose that a nongovernment organization (NGO) is studying the differences in the 50 states in the United States with respect to education and crime. While these two-dimensional data can be easily explored using descriptive statistics and charts, suppose that the NGO is interested in describing each state with a single meta-variable based on the values of the variables Education Spending (dollars per student) and Homicide Rate (homicides per 100,000 people). The file *states* contains information on Education Spending and Homicide Rate for the 50 states in the United States. The values for these variables are standardized so the mean of each variable is zero and the standard deviation of each variable is one. When applying principal component analysis, it is generally recommended to use standardized values of each variable to prevent the varying scales of the variables to distort the process.

We discuss z-scores in Chapter 2 as a means of standardizing variable values.

Applying principal component analysis to this data set, we will describe each of the states with a single meta-variable (called a principal component) instead of the original two variables. Principal component analysis constructs the first principal component (PC_1) of each observation as a weighted linear combination of the original variables. For example, the values of the first principal component PC_1 for Alabama and Alaska can be expressed as

$$\text{Alabama } PC_1 = w_{11} \times \text{Alabama Education Spending} + w_{12} \times \text{Alabama Homicide Rate}$$

$$\text{Alaska } PC_1 = w_{11} \times \text{Alaska Education Spending} + w_{12} \times \text{Alaska Homicide Rate}$$

To understand how principal component analysis determines the values of the weights of the first principal component (w_{11} and w_{12}), consider Figure 6.1 which displays Homicide Rate versus Education Spending for all 50 states. The dispersion of these observations is the "information" contained in these data. For a quantitative variable, variance is a measure of the dispersion of a variable's values. The goal of principal component analysis is to create a new single variable that captures as much information in these two variables as possible. Therefore, principal component analysis sets the values of the weights of the first principal component (w_{11} and w_{12}) to maximize the variance of the resulting PC_1 values. For the data in the file *states*, statistical software supplies first principal component weights of $w_{11} = 0.7071$ and $w_{12} = -0.7071$.

We formally describe the computation of variance in Chapter 2.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6.1 Scatter Chart of Homicide Rate versus Education Spending for the 50 States in the United States

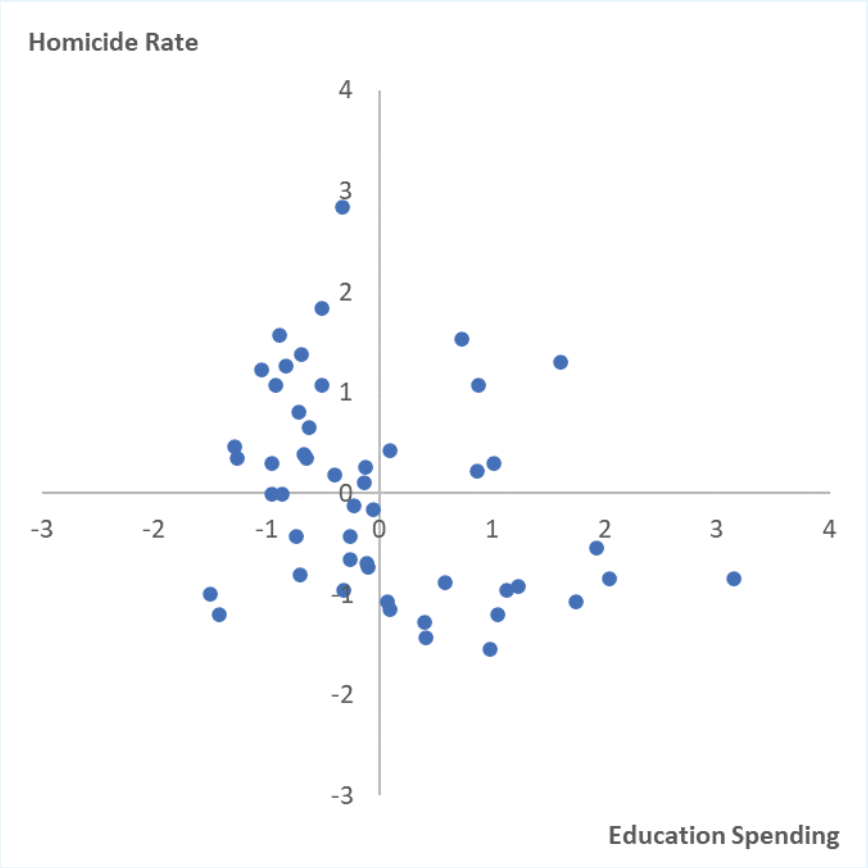


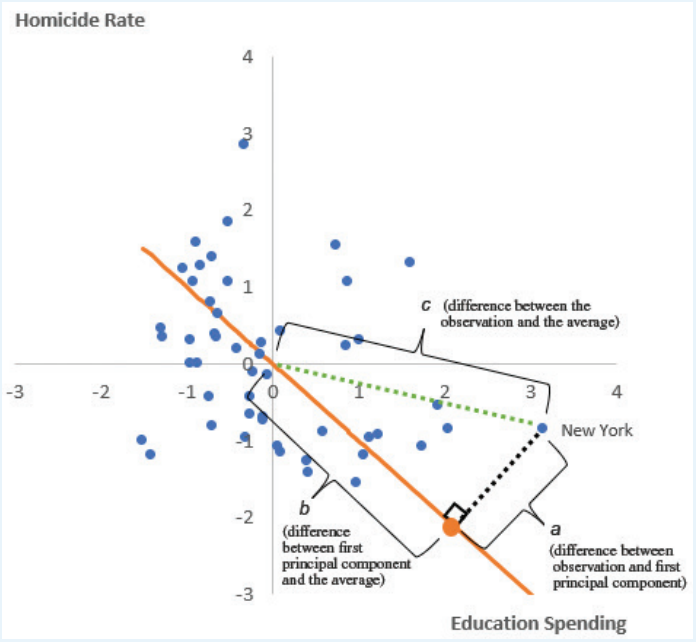
Figure 6.2 facilitates a geometric illustration of the process of determining the weights of the first principal component, w_{11} and w_{12}. These weights determine the slope of the orange line that passes through the origin (0, 0) in Figure 6.2. In turn, the slope of the orange line affects how well the first principal component explains the variance of the observations. We explain the relationship between the slope of the orange line and the observations by considering the New York observation, highlighted in Figure 6.2. The same explanation applies to all other observations.

Segment *a* in Figure 6.2, the length of the black dashed line extending from the New York observation to the point on the orange line, corresponds to the amount of error that results from projecting the two-dimensional New York observation onto a single dimension (the first principal component). Recall because we have standardized the data, the point (0,0) in Figure 6.2 corresponds to the average values of the two variables. Thus, segment *b* in Figure 6.2, the length of the orange line between (0, 0) and the point on the orange line, represents the deviation of the New York observation's first principal component value from the observation corresponding to the average education spending and average homicide rate. Segment *c* is the green dashed line which represents the distance from the New York observation and the observation corresponding to the average education spending and average homicide rate.

The Pythagorean theorem relates the segments *a*, *b*, and *c* in Figure 6.2 through the expression $a^2 + b^2 = c^2$. Note that the length of segment *c* does not depend on the slope of

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6.2 Geometric Interpretation of First Principal Component



the orange line. That is, as you rotate the orange line, the length of the green-dashed line does not change. As a result, using the same calculations over all observations, maximizing the variance of the resulting orange dots (based on projecting each blue dot to the orange line) is the same as minimizing the sum of the resulting values of segment *a*.

Using the weights $w_{11} = 0.7071$ and $w_{12} = -0.7071$, the first principal component values for each of the 50 states can be computed. For example,

$$\begin{aligned}\text{New York } PC_1 &= 0.7071 \times \text{New York Education Spending} \\ &\quad - 0.7071 \times \text{New York Homicide Rate} \\ &= 0.7071 \times 3.1476 - 0.7071 \times -0.8499 = 2.8267\end{aligned}$$

$$\begin{aligned}\text{Louisiana } PC_1 &= 0.7071 \times \text{Louisiana Education Spending} \\ &\quad - 0.7071 \times \text{Louisiana Homicide Rate} \\ &= 0.7071 \times -0.3314 - 0.7071 \times 2.8420 = -2.2439\end{aligned}$$

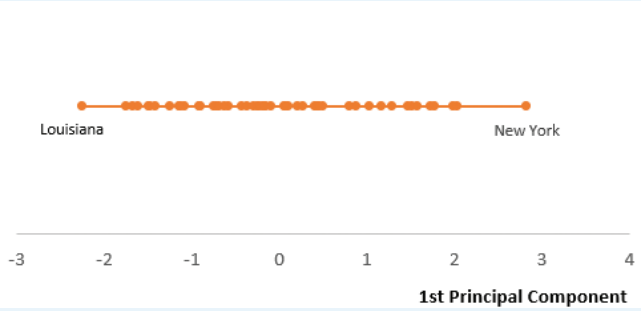
Because the first principal component weight on Education Spending is positive ($w_{11} = 0.7071$) and the first principal component weight on Homicide Rate is negative ($w_{12} = -0.7071$), states with larger PC_1 values will correspond to larger education spending amounts and smaller homicide rates. Conversely, states with smaller PC_1 values will correspond to smaller education spending amounts and larger homicide rates.

Figure 6.3 displays the first principal component values for all 50 states on a line. The display in Figure 6.3 is equivalent to projecting all observations (represented by the blue circles) in Figure 6.2 onto the orange line (as we did with the New York observation) and then rotating this orange line (and the 50 orange points) until it is horizontal.

Over all 50 states, the variance of the PC_1 values is 1.3109. Recall that the variance of each of the original variables, Education Spending and Homicide Rate, was 1.0 because these variables were standardized. Thus, the total variance of the data in the original two

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6.3 First Principal Component Values for All 50 States



variables is 1.0 + 1.0 = 2.0. Therefore, we can say that the first principal component explains 1.3109/2.0 = 65.5% of the variability in the original two variables.

After construction of the first principal component, principal component analysis can be applied again to construct a second principal component. This second principal component is also a weighted linear combination of the original variables. For example, the values of the second principal component PC_2 for Alabama and Alaska can be expressed as

$$\text{Alabama } PC_2 = w_{21} \times \text{Alabama Education Spending} + w_{22} \times \text{Alabama Homicide Rate}$$

$$\text{Alaska } PC_2 = w_{21} \times \text{Alaska Education Spending} + w_{22} \times \text{Alaska Homicide Rate}$$

The weights for the second principal component (w_{21} and w_{22}) are determined so that the second principal component explains as much of the remaining variability in the original variables as possible while having zero correlation with the first principal component. For the data in the file *states*, statistical software supplies the second principal component weights of $w_{21} = 0.7071$ and $w_{22} = 0.7071$. This second principal component explains the remaining 34.5% of the variability in the original two variables. Because the second principal component is constructed in a manner to be independent of the first principal component, the variation in the original two variables that it describes is distinct from the variation described by the first principal component. Therefore, to obtain the total amount of variation explained by the first two principal components, we can add the percentage of variation explained by the first component to the percentage of variation explained by the second component. For the data in the file *states*, the first and second principal components cumulatively explain 65.5% + 34.5% = 100% of the variability in the original two variables. This 100% explanation of variability occurs because describing the data set with two principal components is mathematically equivalent to describing the data set with the two original variables (there is no loss of information because there is no dimension reduction). Figure 6.4 illustrates the first and second principal components. Geometrically, the independence of the first and second principal components results in their corresponding vectors being perpendicular. In effect, principal component analysis has generated an alternative set of axes, one based on PC_1 and PC_2 rather than Educational Spending and Homicide Rate.

Summarizing Protein Consumption for Maillard Riposte

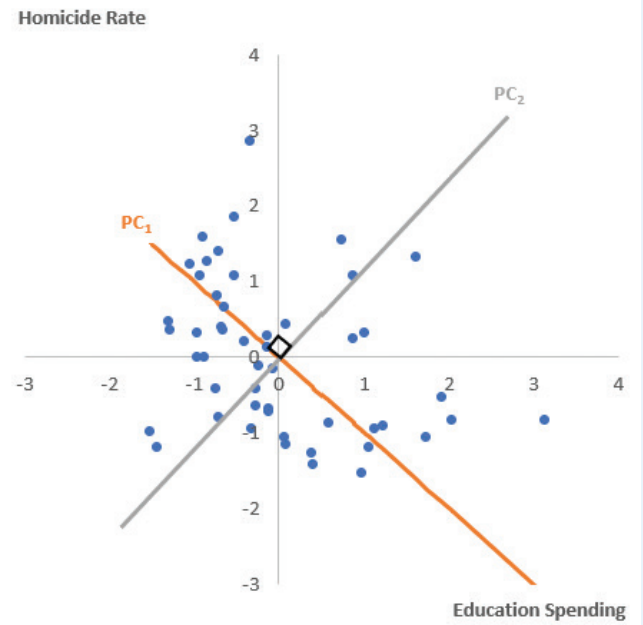
DATA*file*
maillard.csv

Maillard Riposte is a company that develops plant-based substitutes for meat products. As part of its research into expanding into the European market, Maillard Riposte is studying the protein consumption patterns of various countries. The file *maillard* provides the average percentage protein intake from nine major food sources for each of 24 European countries. A portion of these data is shown in Table 6.1.

To simplify the process of identifying similarities and differences in the dietary intake of protein of these 24 countries, we will apply principal component analysis to reduce the number of variables from nine to two. Table 6.2 lists the nine variable weights for both the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6.4 Geometry of First and Second Principal Components



first and second principal components (obtained using statistical software after first standardizing the values of each variable so the mean of each variable is zero and the standard deviation of each variable is one).

A **scree plot** is a data visualization commonly used to evaluate a principal component analysis. For the Maillard Riposte example, Figure 6.5 displays a scree plot of the nine possible principal components. The *i*th point on the blue curve in Figure 6.5 represents the proportion of the variance in the nine original variables explained by the *i*th principal component. For example, the first principal component explains 45.4% of the variability in the original nine variables. The second principal component explains an additional 18.3% of the variability. The *i*th point on the orange curve in Figure 6.5 represents the cumulative proportion of the variance in the nine original variables explained by the first *i* principal components. For example, the cumulative proportion of variance explained by the first two principal components is 0.637 (= 0.454 + 0.183). Thus, together these two principal components explain 63.7% of the variability in the original nine variables.

The scree plot in Figure 6.5 demonstrates that each additional principal component explains a smaller proportion of the remaining variability in the original nine variables.

Table 6.1 Portion of Maillard Riposte Data on Average Percentage Protein Intake

Country	Red Meat	White Meat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fruits & Vegetables
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
Bulgaria	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

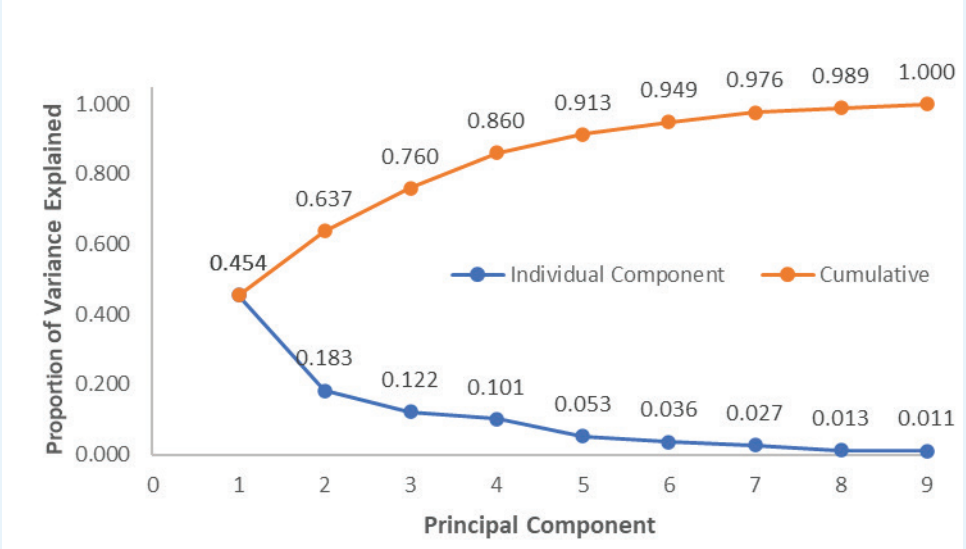
Table 6.2 Weights for First and Second Principal Components

Variable	PC_1 Weight	PC_2 Weight
Red Meat	−0.3116	−0.0376
White Meat	−0.2990	−0.2530
Eggs	−0.4197	−0.0344
Milk	−0.3987	−0.1784
Fish	−0.1475	0.6406
Cereals	0.4309	−0.2419
Starch	−0.2964	0.3373
Nuts	0.4146	0.1658
Fruits & Veggies	0.1058	0.5398

The diminishing value of each additional principal component can guide the analyst in determining how many principal components to select. The number of principal components to select is a subjective decision and depends on how the analyst plans to use the results. For descriptive purposes, fewer principal components may suffice. If the principal components are intended to be used as input variables to predict an outcome variable, more principal components may be desirable. One rule of thumb using a scree plot is to identify the number of principal components beyond which the curves begin to look linear. In Figure 6.5, this rule of thumb would seem to suggest either two or three principal components. In general, the benefit of principal component analysis is that we can use fewer variables (the principal components) to capture a majority of the variability inherent in the original higher-dimensional data.

Loss of variable interpretability is a weakness of dimension reduction methods such as principal component analysis. In the file *maillard*, each country (observation) is described

Figure 6.5 Scree Plot for Principal Component Analysis of Maillard Riposte Data



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 6.3 Variable Values for Austria in Maillard Riposte Data

Variable	Original Units	Standardized Units
Red Meat	8.9	−0.2962
White Meat	14.0	1.7324
Eggs	4.3	1.2618
Milk	19.9	0.3628
Fish	2.1	−0.6537
Cereals	28.0	−0.4544
Starch	3.6	−0.3929
Nuts	1.3	−0.9680
Fruits & Veggies	4.3	−0.0787

by nine variables representing the percentage of protein provided by the respective food sources. These original variables have a clear interpretation. For example, Table 6.3 lists Austria's values for the original nine variables in both the original units and standardized units. Austria's value of Red Meat = 8.9 can be readily interpreted as the average percentage of protein in an Austrian's diet coming from red meat. From the standardized units, we observe that the average percentage of protein in an Austrian's diet coming from red meat is 0.2962 standard deviations below the average of the countries in the data.

Alternatively, we can describe Austria with its first two principal component values (PC_1 and PC_2):

$$\text{Austria } PC_1 = -0.3116 \times -0.2962 + \cdots + 0.1058 \times 0.0787 = -1.48$$

$$\text{Austria } PC_2 = -0.0376 \times -0.2962 + \cdots + 0.5398 \times 0.0787 = -1.09$$

However, Austria's first and second principal component values of −1.48 and −1.09 are not as readily interpretable because each principal component is a combination of all the original nine variables.

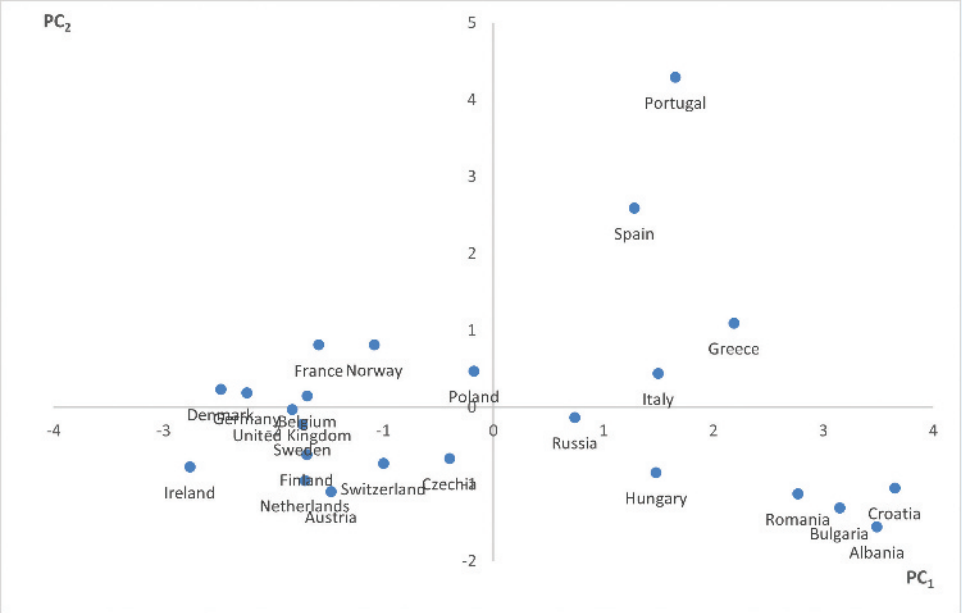
By inspecting the weights of the first and second principal components in Table 6.2, we can perform a "bottom-up" interpretation of these two meta-variables. Examining the weights of PC_1, observe that observations with large PC_1 values will correspond to countries that consume a lot of protein from cereals and nuts (because these two variables have the large, positive PC_1 weights) and/or countries that consume little protein from eggs, milk, red meat, white meat, and starch (because these five variables have very negative PC_1 weights). Conversely, observations with small PC_1 values will correspond to countries that consume a lot of protein from starch, red and white meats, milk, and eggs, and/or very little protein from cereals and nuts. Examining the weights of PC_2, observe that observations with large PC_2 values will correspond to countries that consume a lot of protein from fish and fruits and vegetables (because these two variables have the largest PC_2 weights).

The interpretations of the principal components are subjective and rely upon the analyst's interpretation of the results within the context of the data and the problem.

By examining the scatter chart of the first and second principal component values of the 24 countries in Figure 6.6, we can pair a "top-down" interpretation with our "bottom-up" interpretation. Leveraging contextual knowledge of European countries with the horizontal dispersion of the data displayed in Figure 6.6 reveals that countries with large PC_1 values seem to be less affluent eastern European countries which could explain their consumption of protein from cheaper sources (cereals and nuts). Conversely, countries with smaller PC_1 values tend to be wealthier western European countries who consume protein from more expensive sources (meats and dairy). Focusing on the vertical patterns in the data displayed in Figure 6.6, we observe that Spain and Portugal have large PC_2 values, possibly resulting from a Mediterranean-style diet containing a lot of fish, fruits, and vegetables (such as olives and chickpeas).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6.6 Scatter Chart of First and Second Principal Component Values for 24 Countries



6.2 Cluster Analysis

The goal of cluster analysis is to organize observations into similar groups based on the observed variables. Cluster analysis is commonly used in marketing to divide consumers into different homogeneous groups, a process known as **market segmentation**. Identifying different clusters of consumers allows a firm to tailor marketing strategies for each segment. However, such pattern analysis is not limited to situations in which observations are customers. Cluster analysis is applicable to a wide variety of data. For example, a company may find it insightful to apply clustering to data on its product offerings to identify groups of products that should be managed similarly. Clustering can aid the analysis of firms within the same industry to provide understanding of the market structure. Cluster analysis can be used in finance to group together different stocks with similar characteristics for use in portfolio creation. Cluster analysis can also be used in medical applications such as gene expression analysis where clusters of genes with similar expression patterns are formed to aid disease diagnosis and treatment. Further, analyzing how well individual observations fit into clusters can be used to identify outliers, which in a manufacturing setting may represent quality-control problems and in financial transactions may represent fraudulent activity.

In this section, we present two common clustering methods. First, we consider ***k*-means clustering**, a method which iteratively assigns each observation to one of k clusters in an attempt to achieve clusters that contain observations as similar to each other as possible. Second, we consider agglomerative **hierarchical clustering**, which starts with each observation belonging to its own cluster and then sequentially merges the most similar clusters to create a series of nested clusters. Because both methods rely on measuring the dissimilarity between observations, we first discuss how to calculate distance between observations.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Measuring Distance Between Observations Consisting of Quantitative Variables

DATA*file*
nestegg.csv

Consider a boutique financial advising company called Nest Egg that provides personalized financial advice to its clients. As a basis for developing this tailored advice, Nest Egg would like to segment its clients into groups so that clients within a group are similar with respect to their age and income. The file *nestegg* contains age (measured in years) and income (measured in dollars) for a sample of 44 clients.

When observations consist of quantitative variables, **Euclidean distance** is a common method to measure dissimilarity between observations. Let observations $u = (u_1, u_2, \ldots, u_q)$ and $v = (v_1, v_2, \ldots, v_q)$ each comprise measurements of q variables. The Euclidean distance between observations u and v is

$$d_{uv}^{\text{euclid}} = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_q - v_q)^2}$$

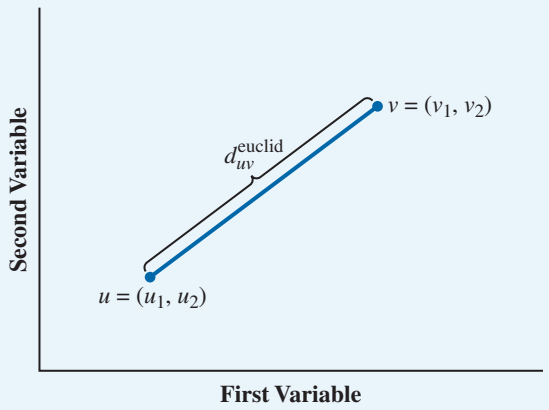
Figure 6.7 depicts Euclidean distance for two observations consisting of two variables ($q = 2$). Euclidean distance becomes smaller as a pair of observations become more similar with respect to their variable values. Euclidean distance is highly influenced by the scale on which variables are measured. For example, consider Nest Egg's task of clustering clients on the basis of the variables Age and Income. Let observation $u = (22, \$25{,}770)$ correspond to a 22-year-old client with an annual income of $25,770 and observation $v = (25, \$25{,}281)$ correspond to a 25-year-old with an annual income of $25,281. As measured by Euclidean distance, the dissimilarity between these two observations is

$$d_{uv}^{\text{euclid}} = \sqrt{(22 - 25)^2 + (25{,}770 - 25{,}281)^2} = \sqrt{9 + 239{,}121} = 489$$

We discuss z-scores in more detail in Chapter 2.

Thus, we see that when using the raw variable values, the amount of dissimilarity between observations is dominated by the Income variable because of the difference in the magnitude of the measurements. Therefore, it is common to standardize the units of each variable j for the observations in a data set. One common standardization technique is to replace the variable values of each observation with the respective z-scores. Suppose that the Age variable has a sample mean of 46 and a sample standard deviation of 13. Also, suppose that the Income variable has a sample mean of 63,699 and sample standard deviation of 20,759.

Figure 6.7 Euclidean Distance



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Then, the Euclidean distance between these two observations based on standardized values is

$$\begin{aligned}\text{(standardized) } d_{uv}^{\text{euclid}} &= \sqrt{\left(\frac{22-46}{13}-\frac{25-46}{13}\right)^2+\left(\frac{25{,}770-63{,}699}{25{,}759}-\frac{25{,}281-63{,}699}{20{,}759}\right)^2}\\ &= \sqrt{(-1.815-(-1.584))^2+(-1.827-(-1.851))^2}\\ &= \sqrt{0.0533+0.0006}=0.23\end{aligned}$$

Based on standardized variable values, we observe that observations u and v are much more different in age than in income. The conversion of the data to z-scores also makes it easier to identify outlier measurements, which can distort the Euclidean distance between observations due to the squaring of the differences in variable values under the square root. Depending on the business goal of the clustering task and the cause of the outlier value, the identification of an outlier may suggest the removal of the corresponding observation, the correction of the outlier value, or the uncovering of an interesting insight corresponding to the outlier observation.

Manhattan distance is a dissimilarity measure that is more robust to outliers than Euclidean distance. The Manhattan distance between observations u and v is

$$d_{uv}^{\text{man}} = |u_1 - v_1| + |u_2 - v_2| + \ldots + |u_q - v_q|$$

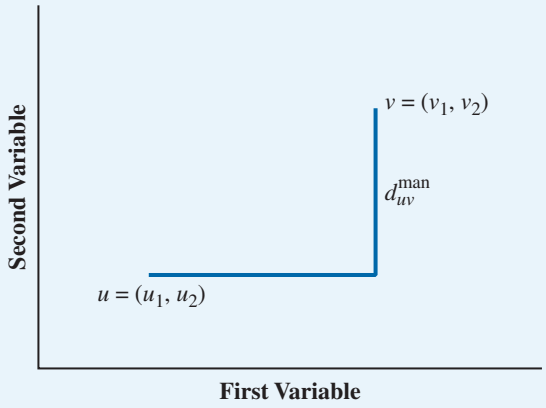
Figure 6.8 depicts the Manhattan distance for two observations consisting of two variables ($q = 2$). From Figure 6.8, we observe that the Manhattan distance between two observations is the sum of the lengths of the perpendicular line segments connecting observations u and v. In contrast to Euclidean distance, which corresponds to the straight-line segment between two observations, Manhattan distance corresponds to the distance as if travelled along rectangular city blocks.

The Manhattan distance between the standardized observations $u = (-1.815, -1.827)$ and $v = (-1.584, -1.851)$ is

$$d_{uv}^{\text{man}} = |(-1.815-(-1.584))| + |-1.584-(-1.851)| = 0.23 + 0.02 = 0.25$$

After conversion to z-scores, unequal weighting of variables can also be considered by multiplying the variables of each observation by a selected set of weights. For instance, after standardizing the units on client observations so that income and age are expressed as their respective z-scores (instead of expressed in absolute values of years and dollars), we can

Figure 6.8 Manhattan Distance



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

multiply the income z-scores by 2 if we wish to treat income with twice the importance of age. In other words, standardizing removes bias due to the difference in measurement units, and variable weighting allows the analyst to introduce any desired bias based on the business context.

Measuring Distance Between Observations Consisting of Categorical Variables

When clustering observations solely on the basis of (unordered) categorical variables, a better measure of similarity between two observations can be achieved by counting the number of variables with matching values. The simplest overlap measure is called the **matching coefficient** and is computed as follows:

Matching Coefficient

$$\frac{\text{number of variables with matching values for observations } u \text{ and } v}{\text{total number of variables}}$$

Subtracting the matching coefficient from 1 results in a distance measure for categorical variables. The **matching distance** between observations u and v (consisting entirely of categorical variables) is

$$\begin{aligned} d_{uv}^{\text{match}} &= 1 - \text{matching coefficient} \\ &= \frac{\text{total number of variables}}{\text{total number of variables}} - \frac{\text{number of variables with matching values}}{\text{total number of variables}} \\ &= \frac{\text{number of variables with mismatching values}}{\text{total number of variables}} \end{aligned}$$

For binary categorical variables, a weakness of the matching coefficient is that if two observations both have a "no" (or "0") value for a variable, this is counted as a sign of similarity between the two observations. However, matching "no" values do not necessarily imply similarity. For example, if a binary categorical variable in a customer survey is Own A Minivan, then a "no" value for two different customers does not mean that they own the same type of car; it means only that neither owns a minivan. The analyst must then determine if "not owning a minivan" constitutes a meaningful notion of similarity between observations in the business context. To avoid misstating similarity due to the absence of a feature, a similarity measure called **Jaccard's coefficient** does not count matching "no" values and is computed as follows:

Jaccard's Coefficient

$$\frac{\text{number of variables with matching "yes" values for observations } u \text{ and } v}{(\text{total number of variables}) - (\text{number of variables with matching "no" values for observations } u \text{ and } v)}$$

Subtracting Jaccard's coefficient from 1 results in the Jaccard distance measure for binary variables. That is, the **Jaccard distance** between observations u and v (consisting entirely of binary variables) is

$$\begin{aligned} d_{uv}^{\text{jac}} &= 1 - \text{Jaccard's coefficient} \\ &= 1 - \frac{\text{number of variables with matching "yes"}}{(\text{total number of variables}) - (\text{number of variables with matching "no"})} \end{aligned}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$$= \frac{(\text{total number of variables}) - (\text{number of variables with matching "no"})}{(\text{total number of variables}) - (\text{number of variables with matching "no"})}$$

$$- \frac{\text{number of variables with matching "yes"}}{(\text{total number of variables}) - (\text{number of variables with matching "no"})}$$

$$= \frac{\text{number of variables with mismatching values}}{\text{total number of variables} - \text{number of variables with matching "no"}}$$

Suppose Nest Egg has surveyed its clients and collected information on four binary categorical variables listed below.

Masters	whether client has a master's degree or beyond
Married	whether client is married
Family	whether client has dependent children
Mortgage	whether client has a mortgage on a home

For five of Nest Egg's clients, Table 6.4 contains observations of the binary categorical variables Masters, Married, Family, and Mortgage and the dissimilarity matrixes based on the matching distance and Jaccard's distance, respectively. Based on the matching distance, Observation 1 and Observation 4 are more similar (0.25) than Observation 2 and Observation 3 (0.5) because 3 out of 4 variable values match between Observation 1 and Observation 4 versus just 2 matching values out of 4 for Observation 2 and Observation 3. However, based on Jaccard's distance, Observation 1 and Observation 4 are equally similar (0.5) as Observation 2 and Observation 3 (0.5) because Jaccard's coefficient discards the matching "no" values for the Family and Mortgage variables for Observation 1 and Observation 4. In the context of this

Table 6.4 Comparison of Distance Matrixes for Observations with Binary Variables

Observation	Masters	Married	Family	Mortgage
1	1	0	0	0
2	0	1	1	1
3	1	1	1	0
4	1	1	0	0
5	1	1	0	0

Matrix of Matching Distances

Observation	1	2	3	4	5
1	0.000				
2	1.000	0.000			
3	0.500	0.500	0.000		
4	0.250	0.750	0.250	0.000	
5	0.250	0.750	0.250	0.000	0.000

Matrix of Jaccard Distances

Observation	1	2	3	4	5
1	0.000				
2	1.000	0.000			
3	0.667	0.500	0.000		
4	0.500	0.750	0.333	0.000	
5	0.500	0.750	0.333	0.000	0.000

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

example, choice of the matching distance or Jaccard's distance depends on whether Nest Egg believes that matching "no" entries imply similarity. That is, Nest Egg must gauge whether meaningful similarity is implied if a pair of observations do not have an advanced degree, not married, do not have dependent children, or do not have a mortgage. For these data, it seems plausible that matching distance is appropriate.

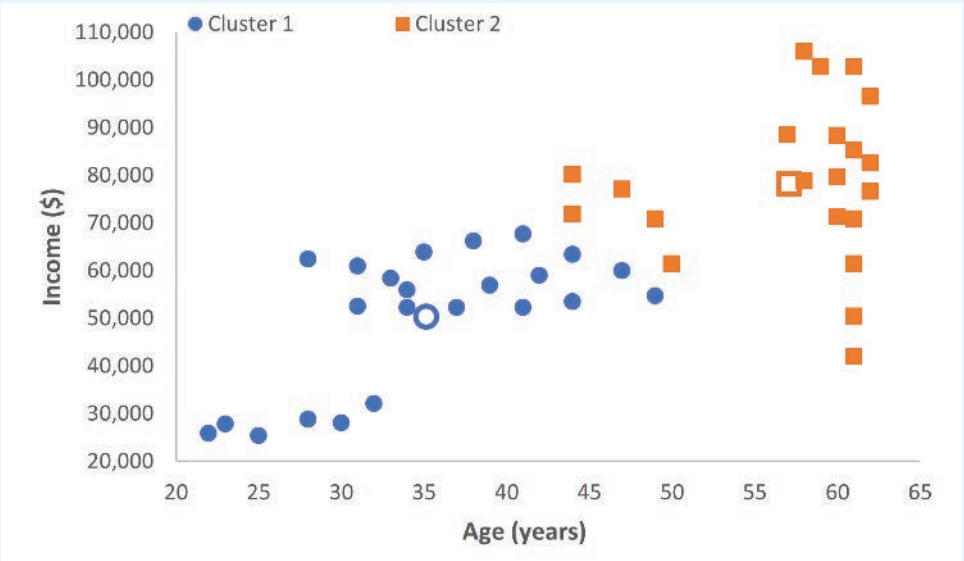
k-Means Clustering

When considering observations consisting entirely of quantitative variables, an approach called *k*-means clustering is commonly used to organize observations into similar groups. In *k*-means clustering, the analyst must specify the number of clusters, *k*. Given a value of *k*, the *k*-means algorithm begins by assigning each observation to one of the *k* clusters. After all observations have been assigned to a cluster, the cluster centroids for each cluster are computed by determining the average value of each variable for the observations in a respective cluster. The *k* cluster centroids are the "means" of *k*-means clustering. Using the updated cluster centroids, all observations are reassigned to the cluster with the closest centroid (where Euclidean distance is the standard metric). If any observation is assigned to a different cluster, this will change the cluster centroids, so the algorithm repeats this process (calculate cluster centroid, assign each observation to the cluster with nearest centroid) until there is no change in the clusters or a specified maximum number of iterations is reached.

DATA*file*
nestegg.csv

If the number of clusters, *k*, is not clearly established by the context of the business problem, the *k*-means clustering algorithm can be repeated for several values of *k* to assess the impact on the resulting set of clusters. Figures 6.9, 6.10, and 6.11 depict the resulting clusters when applying *k*-means clustering for $k = 2$, 3, and 4, respectively, to Nest Egg's client data in the file *nestegg*. In each of these figures, the centroids of the respective clusters are represented by hollow shapes (e.g., in Figure 6.9, ☐ and ○). While Figures 6.9, 6.10, and 6.11 are plotted in the original scale of the variables Age and Income, the clustering was based on the standardized scale of these variables.

Figure 6.9 Grouping Nest Egg Clients by Age and Income Using *k*-Means Clustering with $k = 2$



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6.10 Grouping Nest Egg Clients by Age and Income Using *k*-Means Clustering with $k = 3$

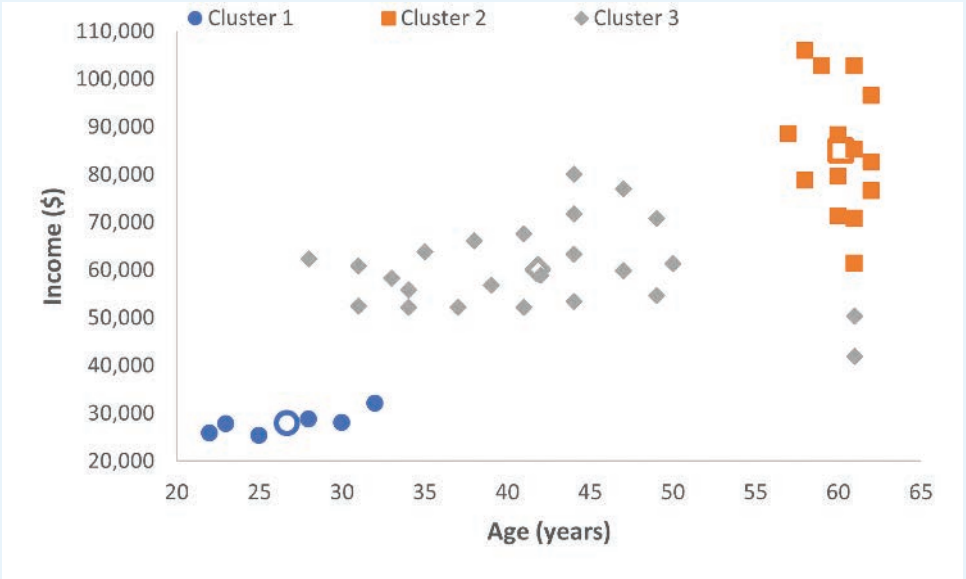


Figure 6.11 Grouping Nest Egg Clients by Age and Income Using *k*-Means Clustering with $k = 4$



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Once we have constructed one or more sets of clusters, we would like to assess the clusters. As an unsupervised learning technique, cluster analysis is not guided by any explicit measure of accuracy, and thus the notion of a "good" clustering is subjective. Nonetheless, we describe three methods for assessing a set of clusters: (1) cluster cohesion and cluster separation, (2) cluster interpretability, and (3) cluster stability.

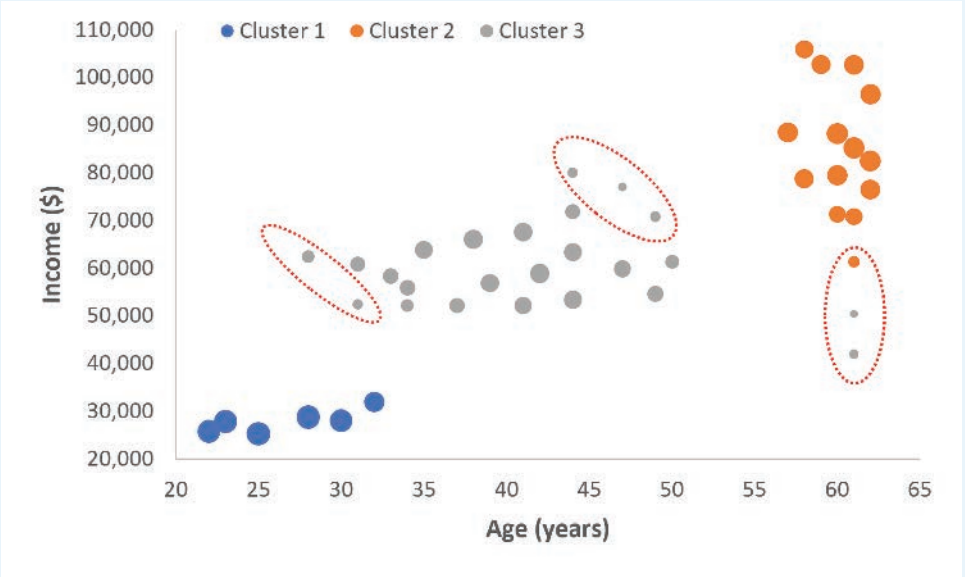
Clustering should result in observations being grouped so that they are more similar to observations in the same cluster and less similar to observations in other clusters. **Cluster cohesion** relates to the distance between observations within the same cluster. **Cluster separation** relates to the distance between observations in different clusters.

Silhouette is a measure that combines the notions of cluster cohesion and cluster separation. To measure cluster cohesion, define $a(i)$ as the average distance from observation i to the other observations in the same cluster as observation i. To measure cluster separation, define $b(i)$ as the average distance from observation i to the observations in the closest neighboring cluster to observation i. The silhouette value for an observation i assigned to a cluster with more than one observation is then calculated as

$$\text{Silhouette of observation } i = s(i) = \frac{b(i) - a(i)}{max\,\{a(i), b(i)\}}$$

For an observation i that is the only observation in its cluster, $s(i)$ is set to 0 because the value of $a(i)$ is not well-defined in this case. The values of $s(i)$ range from -1 to 1. A value of $s(i)$ close to 1 means observation i is much closer to the observations in its cluster than to observations in the next closest cluster; in other words, $a(i)$ is much smaller than $b(i)$. Therefore, a larger value of $s(i)$ represents a better match between observation i and its cluster. A value of $s(i)$ close to 0 means observation i is on the border of two clusters as it is equidistant from observations in its current cluster and a neighboring cluster. Finally, as $s(i)$ becomes negative, it means that observation i is closer to observations in a cluster other than its own. Figure 6.12 displays the same set of three clusters as Figure 6.10, but the size

Figure 6.12 Illustrating Silhouette Values of Observations Using Marker Size in the *k*-Means Clustering of Nest Egg Clients with $k = 3$



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 6.5 Average Silhouette Values for *k*-Means Clustering

Number of Clusters (*k*)	Average Silhouette Value
2	0.478
3	0.507
4	0.456

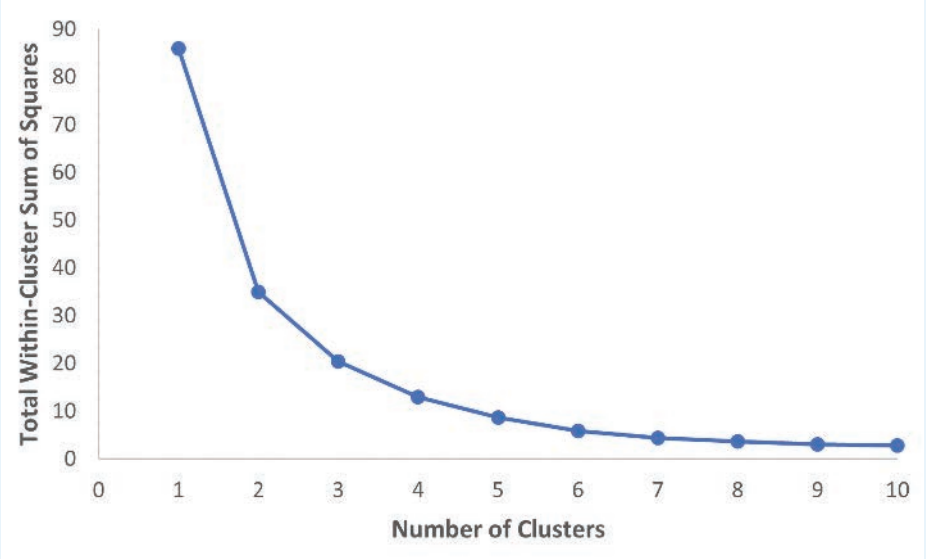
of each data point is proportional to the corresponding observation's silhouette value. From Figure 6.12, we observe that the observations with the smallest silhouette values (circled in red) occur on the edges of a cluster that is relatively close to a different cluster, while the observations with the largest silhouette values occur in the heart of a cluster or in a cluster that is not relatively close to another cluster.

For a given set of clusters, averaging the silhouette values of all the observations provides a measure of how well the entire data set has been clustered. Table 6.5 lists average silhouette values for the sets of clusters in Figures 6.9, 6.10, and 6.11, respectively. For these 44 clients, the set of 3 clusters results in the largest average silhouette value. However, recall that the goal of clustering is not necessarily to maximize the average silhouette value. This is just one measure to aid the analyst in the cluster assessment.

The **elbow method** is another common approach to assess clustering based on cluster cohesion. Using the total within-cluster sum of squared distances from the observations to their assigned cluster centroid as a measure of cluster cohesion, the elbow method compares the effect of varying the number of clusters.

Figure 6.13 contains the elbow chart that visualizes the effect of varying the number of clusters ($k = 1, \ldots 10$) for the *k*-means clustering on the Nest Egg data. We observe from Figure 6.13 that as the number of clusters increases, the total within-cluster sum of squared distances decreases. If we allow the number of clusters to be equal to the total number of

Figure 6.13 Elbow Chart for Nest Egg Data



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

observations, then the within-cluster sum of squared distances is zero. Of course, placing each observation in its own cluster provides no insight into the similarity between observations, so minimizing the total within-cluster sum of squared distance is not the goal. From Figure 6.13, we also observe that the marginal decrease in the total within-cluster sum of squared distances diminishes as k gets larger. To identify the elbow point on the plot, look for the point beyond which the curve begins to look linear. In Figure 6.13, we see that the most promising values of k seem to be 3 or 4, suggesting that the sets of clusters for $k = 3$ and $k = 4$ should be examined more closely.

Another consideration when assessing clusters is **cluster interpretability**. In addition to measures of how cohesive and separated each cluster is, an analyst may also be interested in how much interpretable insight the clusters provide. A common use of clustering is to use cluster centroids as prototypes to summarize observations that belong to the respective clusters. Therefore, obtaining a set of clusters which provide intuitive interpretations within the context of the data's application is often useful.

For the set of three clusters in Figure 6.10, Cluster 1 is characterized by relatively younger, lower-income clients (Cluster 1's centroid is at [26.7, $27,936]). Cluster 2 is characterized by relatively older, higher-income clients (Cluster 2's centroid is at [60.1, $85,049]). Cluster 3 is characterized by moderately aged, middle-income clients (Cluster 3's centroid is at [41.8, $60,185]). Analogous interpretations of the clustering in Figure 6.9 and the clustering in Figure 6.11 can be conducted. Pairing how well these interpretations fit Nest Egg's qualitative perception of its client base with measures of cluster separation and cohesion, a preferred set of clusters can be identified. In this case, it appears that the set of three clusters may be preferred.

For large data sets, another method for assessing a set of clusters is to evaluate **cluster stability**. Cluster stability refers to how consistent the set of clusters obtained are when there are slight changes in the data. It is undesirable for the set of clusters that a clustering method obtains to change dramatically when there are small changes in the observations used as input. This instability casts doubt on the validity of the clusters. Therefore, for data sets with many observations, the robustness of the clusters can be evaluated with the following process:

Step 1. Remove a small percentage of observations from the data.
Step 2. Cluster the remaining observations.
Step 3. Assign removed observations to the most similar cluster from step 2.
Step 4. Compare clusters obtained in step 3 to clusters formed by applying the clustering method to all observations.

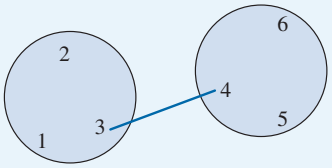
Hierarchical Clustering and Measuring Dissimilarity Between Clusters

An alternative to partitioning observations with the k-means approach is an agglomerative hierarchical clustering approach that starts with each observation in its own cluster and then iteratively combines the two clusters that are the least dissimilar (most similar) into a single cluster. Each iteration corresponds to an increased level of aggregation by decreasing the number of distinct clusters. Hierarchical clustering determines the dissimilarity of two clusters by considering the distance between the observations in first cluster and the observations in the second cluster. Given a way to measure distance between pairs of individual observations (e.g., Euclidean distance, Manhattan distance, matching distance, Jaccard distance), there are several agglomeration methods for comparing observations in two clusters to obtain a cluster dissimilarity measure. Using Euclidean distance to illustrate, Figure 6.14 provides a two-dimensional depiction of four agglomeration methods we will discuss.

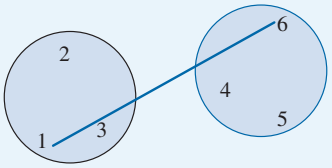
When using the **single linkage** agglomeration method, the dissimilarity between two clusters is defined by the distance between the pair of observations (one from each cluster) that are the most similar. Thus, single linkage will consider two clusters to be similar if an observation in one of the clusters is similar to at least one observation in the other cluster. However, a cluster formed by merging two clusters that are similar with respect to single

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

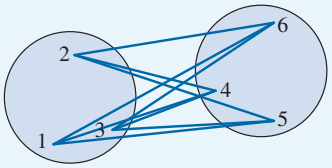
Figure 6.14 Measuring Dissimilarity Between Clusters



Single Linkage, $d_{3,4}$

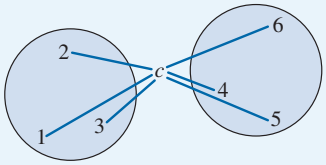


Complete Linkage, $d_{1,6}$



Group Average Linkage,

$$\frac{d_{1,4}+d_{1,5}+d_{1,6}+d_{2,4}+d_{2,5}+d_{2,6}+d_{3,4}+d_{3,5}+d_{3,6}}{9}$$



Ward's Linkage,

$$d_{1c}^2+d_{2c}^2+d_{3c}^2+d_{4c}^2+d_{5c}^2+d_{6c}^2$$

linkage may also consist of pairs of observations that are very different. The reason is that there is no consideration of how different an observation may be from other observations in a cluster as long as it is similar to at least one observation in that cluster. Thus, in two dimensions (variables), single linkage clustering can result in long, elongated clusters rather than compact, circular clusters.

The **complete linkage** agglomeration method defines the dissimilarity between two clusters as the distance between the pair of observations (one from each cluster) that are the most different. Thus, complete linkage will consider two clusters to be similar if their most-different pair of observations are similar. This method produces clusters such that all member observations of a cluster are relatively similar to each other. The clusters produced by complete linkage have approximately equal diameters. However, complete linkage clustering can be distorted by outlier observations.

The single linkage and complete linkage methods both define between-cluster dissimilarity based on a single pair of observations. In contrast, the **group average linkage** agglomeration method defines the dissimilarity between two clusters to be the average distance computed over *all* pairs of observations between the two clusters. If Cluster 1 consists of n_1 observations and Cluster 2 consists of n_2 observations, the dissimilarity of these clusters would be the average of $n_1 \times n_2$ distances. This method produces clusters that are less dominated by the dissimilarity between single pairs of observations.

Ward's linkage agglomeration method for merging clusters is based on the notion that representing a cluster with its centroid can be viewed as a loss of information in the sense that the individual differences in the observations within the cluster are not captured by the cluster centroid. For a pair of clusters under consideration for

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

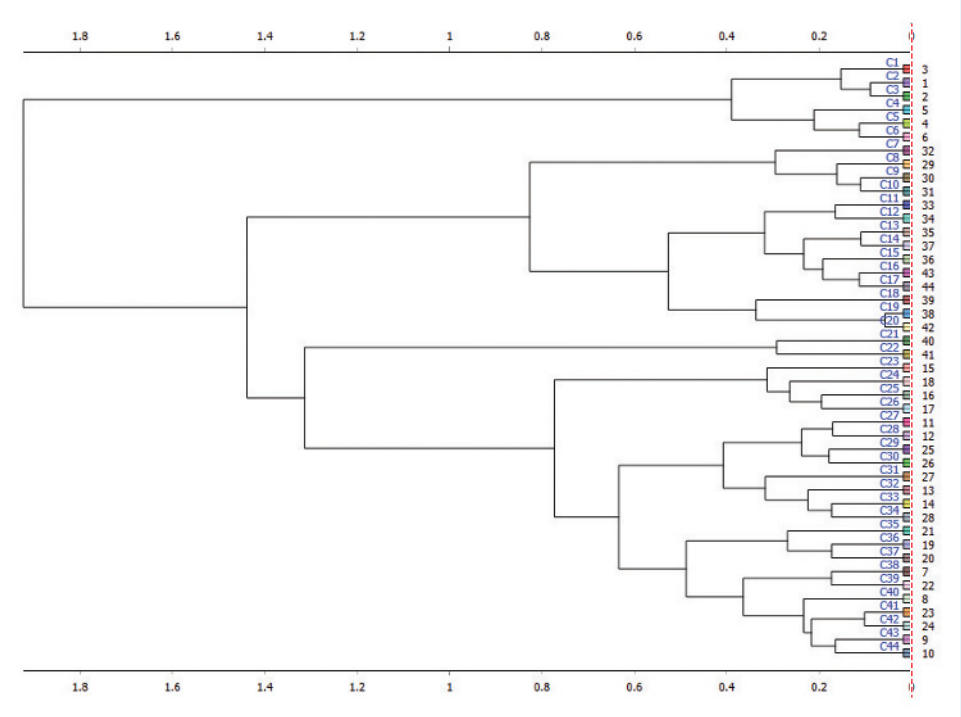
aggregation, Ward's linkage computes the centroid of the resulting merged cluster and then calculates the sum of squared distances between this centroid and each observation in the union of the two clusters. At each iteration, Ward's linkage merges the pair of clusters with the smallest value of this dissimilarity measure. As a result, hierarchical clustering using Ward's linkage results in a sequence of aggregated clusters that minimizes the loss of information between the individual observation level and the cluster centroid level.

Returning to the Nest Egg example, we demonstrate agglomerative hierarchical clustering on the age and income data for 44 Nest Egg clients in the file *nestegg*. Figure 6.15 depicts a **dendrogram** that visually summarizes the output from agglomerative hierarchical clustering using Euclidean distance to measure dissimilarity between observations and the group average linkage agglomeration method to measure dissimilarity between clusters. A dendrogram is a chart that depicts the set of nested clusters resulting at each step of aggregation. In this dendrogram, the vertical axis of the dendrogram lists the 44 observation indexes. Reading from right to left, the horizontal axis of the dendrogram represents the dissimilarity (distance) resulting from a merger of two different groups of observations. Because agglomerative hierarchical clustering begins with each observation in its own cluster, this clustering procedure begins with 44 singleton clusters for the Nest Egg data as shown on the far right of Figure 6.15. Each of these 44 clusters is represented by a horizontal line extending from the observation index. When two clusters are merged, a vertical line links the two horizontal lines corresponding to the clusters. The placement of this vertical line is determined by the degree of dissimilarity (distance) in the two merged clusters.

DATA*file*
nestegg.csv

Dendrograms may also be oriented so that the horizontal axis lists the observation indexes and the vertical axis represents the dissimilarity (distance) resulting from the merger of clusters.

Figure 6.15 Dendrogram for Nest Egg Data Using Euclidean Distance and Group Average Linkage



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

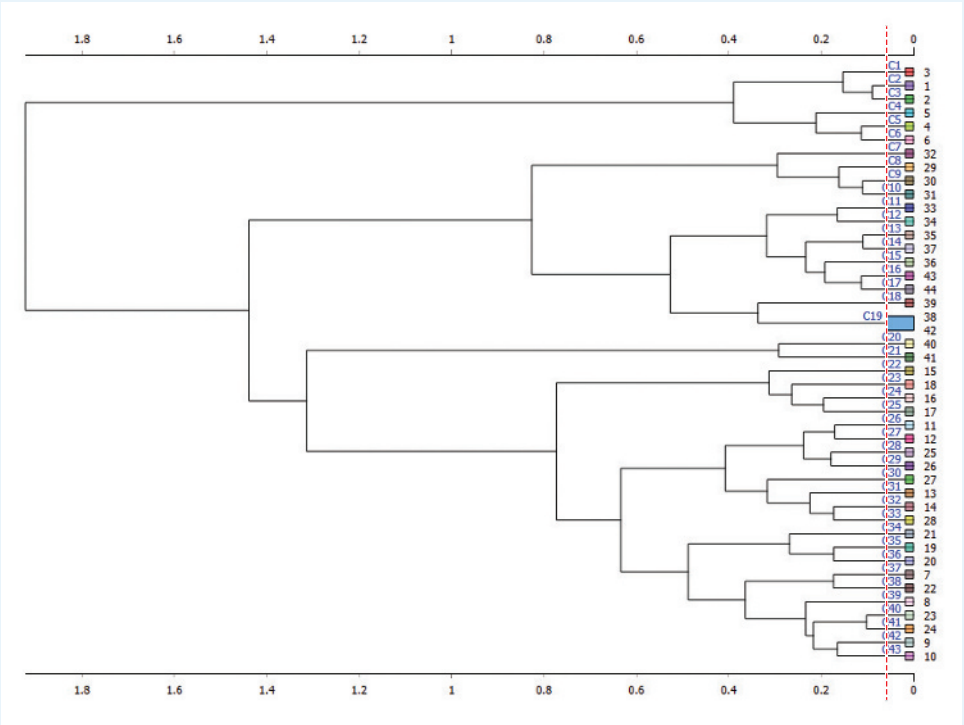
By drawing a dashed vertical line from the top to the bottom of the dendrogram, we can obtain the number of clusters at a specified level of aggregation. For example, in Figure 6.15, the dashed vertical line drawn at a distance value of zero on the horizontal axis intersects with 44 horizontal lines representing that the beginning of an agglomerative hierarchical process there is no dissimilarity within any of the clusters because each cluster contains a single observation.

Increasing levels of aggregation can be observed by sliding the dashed vertical line farther to the left. For example, Figure 6.16 shows that the first merger occurs when the cluster containing observation 38 is merged with the cluster containing observation 42. This merger occurs at a value of 0.0806 on the horizontal axis and that represents the amount of dissimilarity introduced by merging these two clusters.

By sliding the dashed vertical line left or right along the vertical axis of a dendrogram and observing the intersection of the dashed vertical line with the horizontal branches, an analyst can extract varying number of clusters. Figure 6.17 displays the dendrogram with a vertical cut corresponding to four clusters along with the scatter chart illustrating these four clusters according to their age and income values. Similarly, Figure 6.18 displays the dendrogram with a vertical cut corresponding to three clusters along with the scatter chart illustrating these three clusters according to their age and income values. The set of two clusters obtained by agglomerative hierarchical clustering based on Euclidean distance and group average linkage on the Nest Egg data are illustrated by the dendrogram and scatter chart in Figure 6.19.

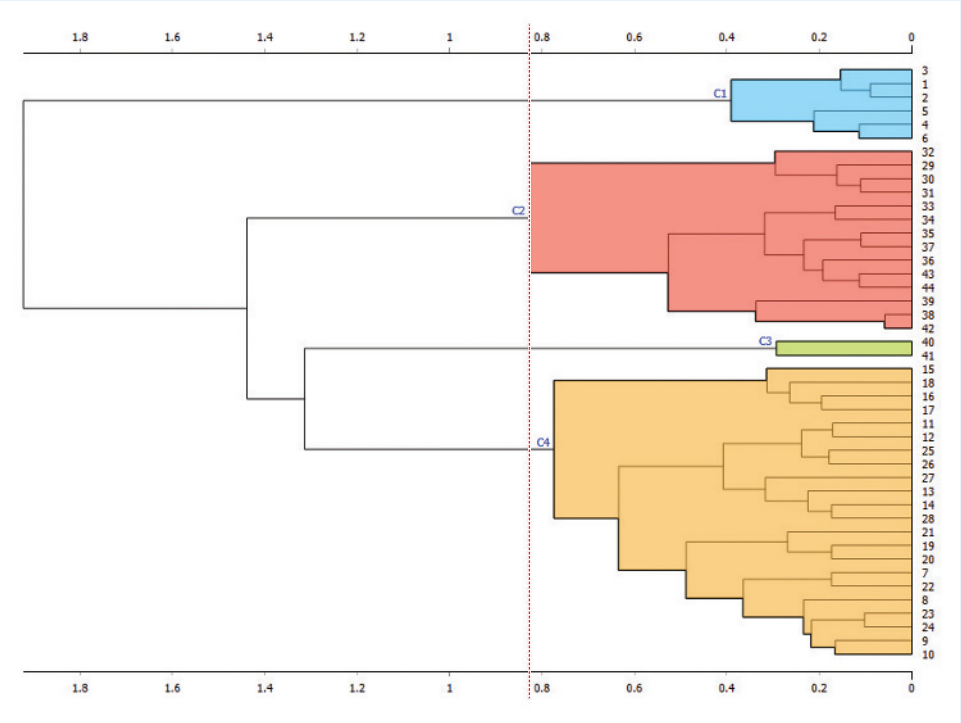
Comparing Figure 6.17 to the k-means clustering with $k = 4$ in Figure 6.11, we observe that k-means clustering and hierarchical clustering result in different clusters.

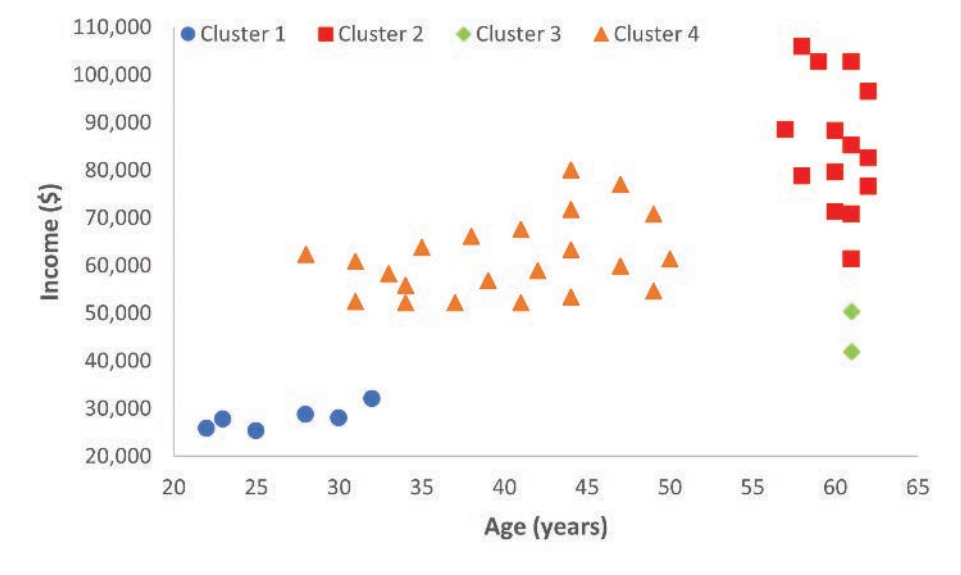
Figure 6.16 Dendrogram on Nest Egg Data with Vertical Cut Depicting the First Merger of a Pair of Clusters



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6.17 Dendrogram and Scatter Chart of Four Clusters on Nest Egg Data

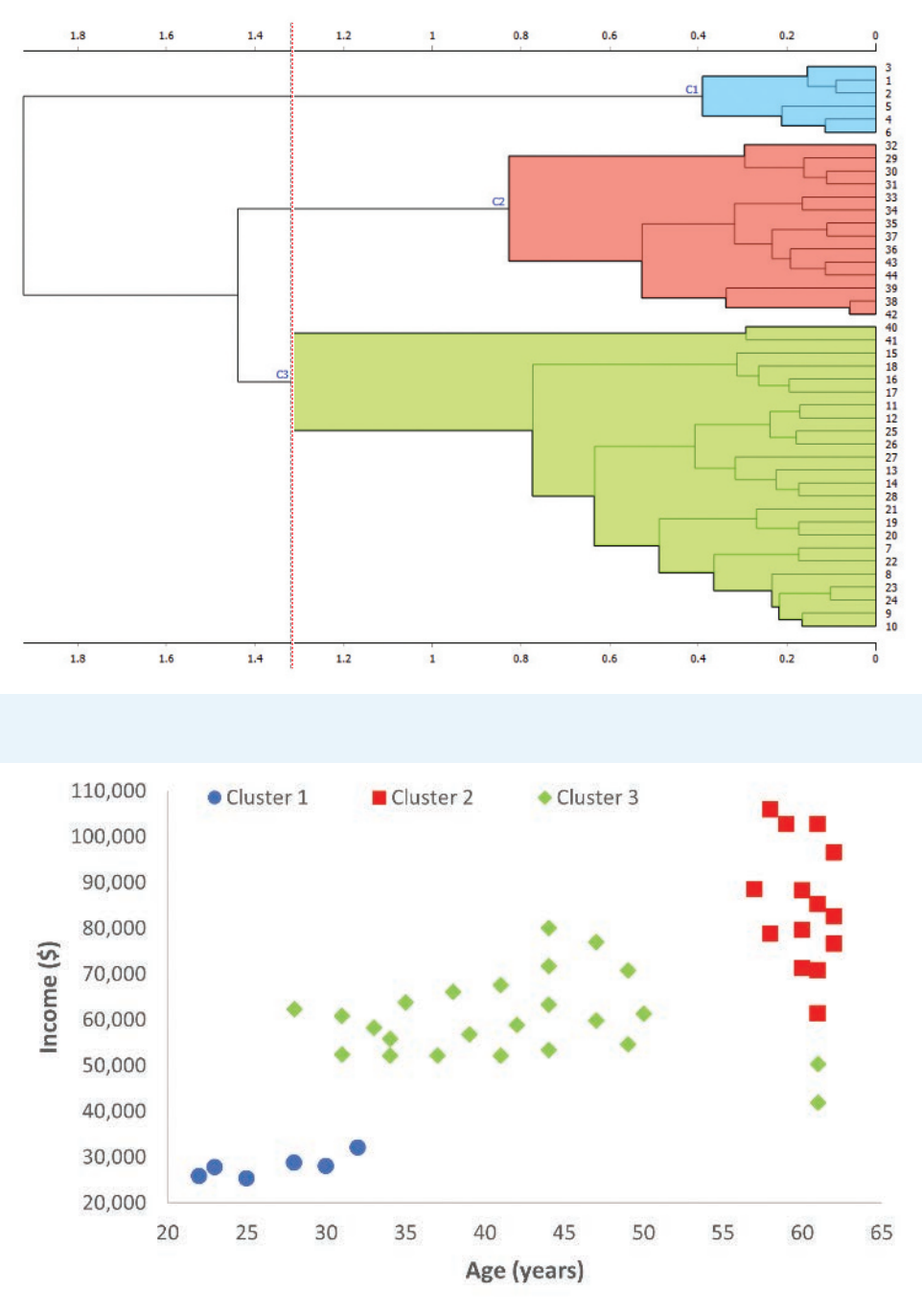




A comparison of Figure 6.18 to the *k*-means clustering with $k = 3$ in Figure 6.10 reveals that for these data, *k*-means clustering and hierarchical clustering obtain the same set of three clusters. Comparing Figure 6.19 to the *k*-means clustering with $k = 2$ in Figure 6.9, we again observe that *k*-means clustering and hierarchical clustering result in different clusters.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

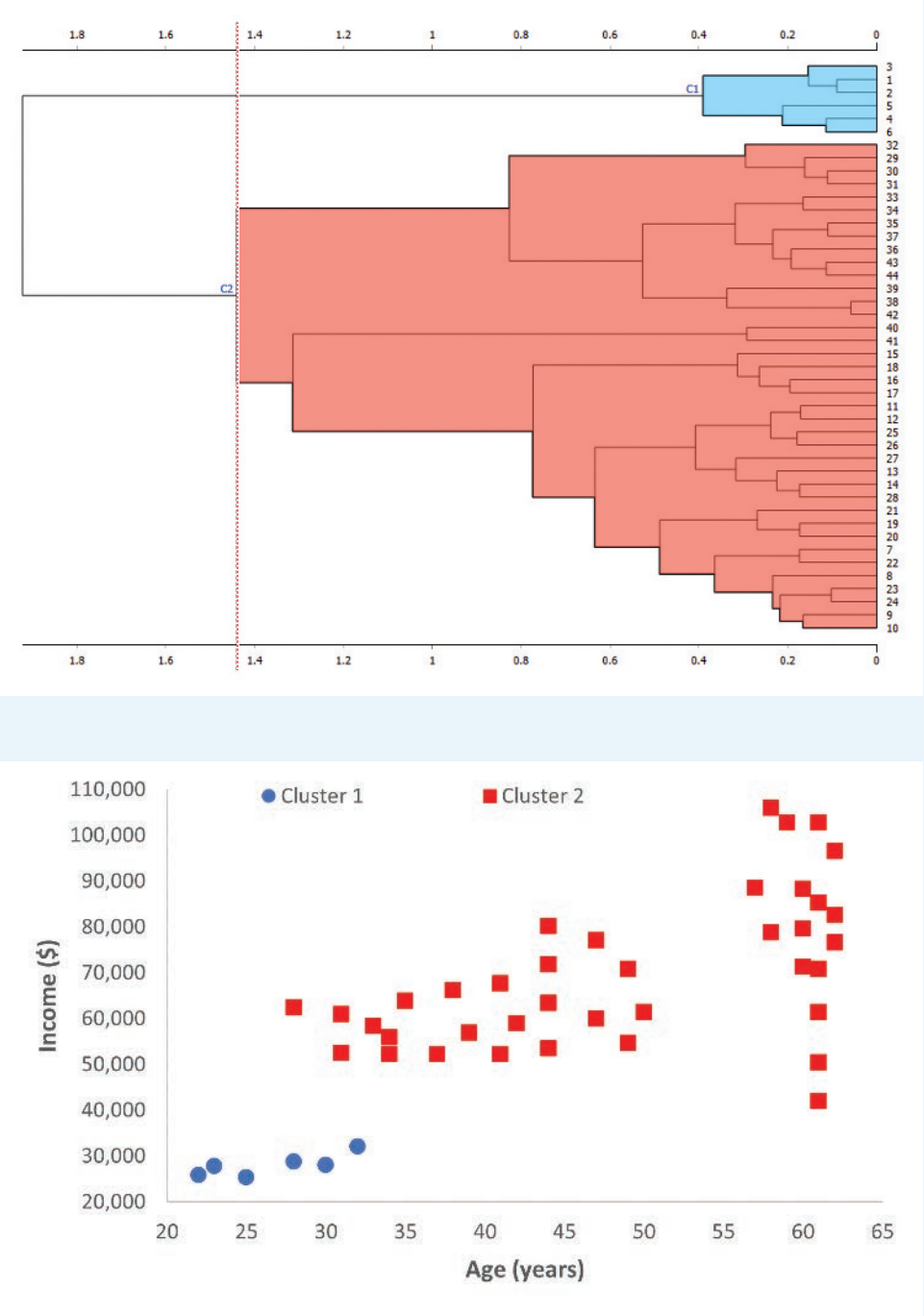
Figure 6.18 Dendrogram and Scatter Chart of Three Clusters on Nest Egg Data



Even though both methods are using Euclidean distance for these data, they may obtain different clusters because they use different processes to form the clusters. In *k*-means clustering, there are always *k* clusters, but for agglomerative hierarchical clustering a set of *k* clusters is obtained by sequentially merging individual clusters until obtaining a set of *k* clusters. Because existing clusters are sequentially merged in the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6.19 Dendrogram and Scatter Chart of Two Clusters on Nest Egg Data



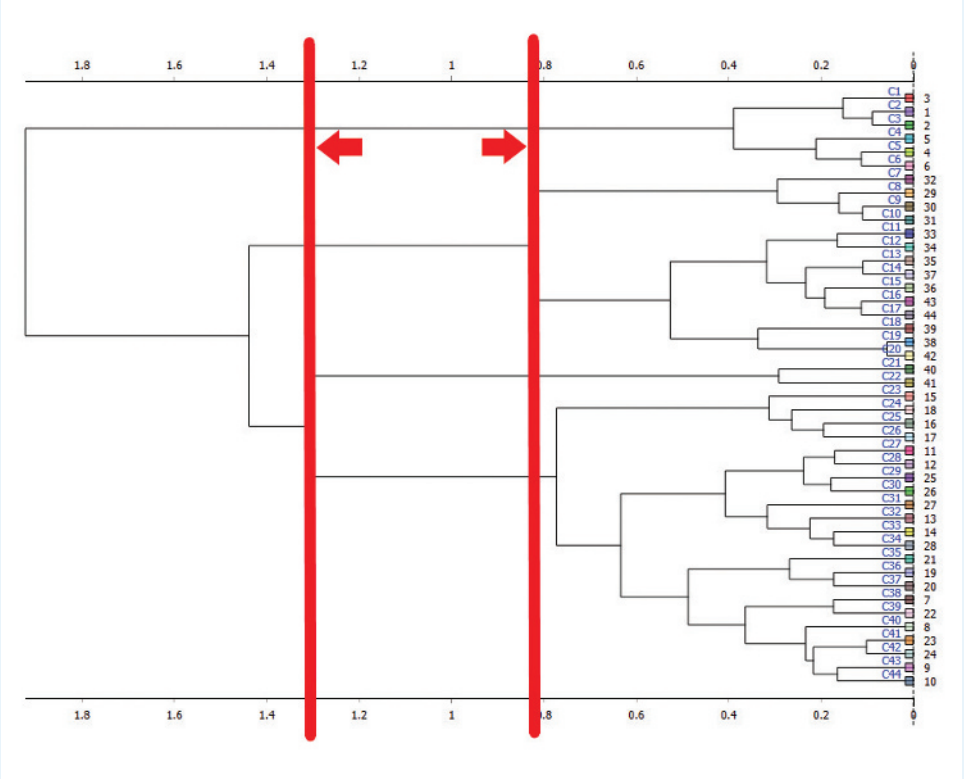
hierarchical clustering process if a pair of observations is in the same group in a set of k clusters then this pair of observations will also be in the same group in a set of $k - 1$ clusters. However, in k-means clustering, a pair of observations in the same group in a set of k clusters may not be in the same group when the value of k is decremented to find a set of $k - 1$ clusters.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To assess a set of clusters obtained via hierarchical clustering, the concepts of cluster cohesion and cluster separation, cluster interpretability, and cluster stability can be applied. A dendrogram provides some visual clues about what may be an appropriate level of aggregation. In Figure 6.20, the right-most red vertical line represents the stage in the hierarchical clustering process with a set of four clusters. Similarly, the left-most red vertical line in Figure 6.20 represents the stage in the hierarchical clustering process with a set of three clusters. Thus, the horizontal distance between these two red vertical lines corresponds to the "cost" of merging a pair of clusters from the set of four clusters to obtain set of three clusters. This cost is measured in terms of the increase in the dissimilarity within clusters. So, horizontally elongated portions of a dendrogram represent mergers of more dissimilar clusters, and horizontally compact portions of a dendrogram represent mergers of more similar clusters. Using the horizontal elongation and compactness of the dendrogram in Figure 6.20 as a guide, it appears that a set of four clusters is promising. However, depending on the number of client profiles that Nest Egg may prefer, it appears that grouping the clients into a range of two to seven clusters may be reasonable.

A dendrogram also provides visual indication of the cluster separation of individual clusters. Specifically, an individual cluster's separation can be gauged by the horizontal distance between the point at which cluster is originally formed and the point at which it is merged with another cluster. In Figure 6.19, the cluster containing the observation set {3, 1, 2, 5, 4, 6} (shaded in blue) is formed relatively early in the process (when there are 10 clusters) and remains its own cluster until hierarchical clustering groups the entire collection of observations into one single cluster of 44 observations.

Figure 6.20 Interpreting Distance Between Successive Mergers in a Dendrogram



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Hierarchical Clustering versus *k*-Means Clustering

Hierarchical clustering is a good choice in situations in which you want to easily examine solutions with a wide range of clusters. Hierarchical clusters are also convenient if you want to observe how clusters are nested. However, hierarchical clustering can be very sensitive to outliers, and clusters may change dramatically if observations are eliminated from, or added to, the data set. Hierarchical clustering may be a less appropriate option as the number of the observations in the data set grows large because hierarchical clustering is relatively computationally expensive.

The *k*-means approach is a good option for clustering data on the basis of numerical variables, and is computationally efficient enough to handle a large number of observations. Recall that *k*-means clustering partitions the observations, which is appropriate if you are trying to summarize the data with *k* "average" observations that describe the data with the minimum amount of error. However, *k*-means clustering is generally not appropriate for categorical or ordinal data, for which an "average" is not meaningful.

For both hierarchical and *k*-means clustering, the selection of variables on which to base the clustering process is critical. Clustering should be based on a parsimonious set of variables, determined through a combination of context knowledge and experimentation with various variable combinations, that reveal interesting patterns in the data. As the number of variables used in a clustering approach increases, all observations tend to become equidistant from each other and the relative differences in the distances between pairs of observations tend to get smaller.

As a final note, whichever clustering method is applied to data, it will generate a set of clusters even if the data does not contain patterns amenable to grouping observations. It is up to the analyst to determine if the clusters obtained are meaningful and provide insight, rather than just relatively arbitrary groupings of observations.

Notes + Comments

1. Clustering observations based on both quantitative and categorical variables (mixed data) can be challenging because distance measures appropriate for quantitative variables (e.g., Euclidean and Manhattan) are not appropriate for categorical variables. For non-ordinal categorical variables, the standard approach is to binary encode the variables and use a distance measure such as matching distance or Jaccard distance. Ordinal categorical variables introduce additional obstacles because numerically encoding ordered levels (e.g., 1, 2, 3, . . .) implicitly requires the analyst to define a distance between the levels.

   One approach to clustering mixed data is to convert all variables into binary variables. Converting quantitative and ordinal variables into binary categorical variables results in a loss of information, but allows the use of a single appropriate measure (e.g., matching distance or Jaccard distance).

   Another approach to clustering mixed data is to use Gower's distance measure. For each variable, Gower's distance measure computes the distance between variable values for a pair of observations using a distance measure appropriate for the variable type and scales the distance between [0,1]. Then the scaled distance measures obtained for each variable are combined in a weighted average to obtain a single distance measure for a pair of observations. The statistical software package R has functionality that implements Gower's distance.
2. Centroid linkage, median linkage, and McQuitty's linkage are three additional linkage methods commonly implemented in statistical software. However, when using median linkage or centroid linkage to determine a pair of clusters to merge, the total dissimilarity in a set of $k - 1$ clusters can actually be less than a set of k clusters. This counterintuitive non-monotonic behavior is a drawback of using median linkage and centroid linkage. McQuitty's linkage is similar to group average linkage, but weights the observations differently.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6.3 Association Rules

In marketing, analyzing consumer behavior can lead to insights regarding the placement and promotion of products. Specifically, marketers are interested in examining transaction data on customer purchases to identify the products commonly purchased together. Bar-code scanners facilitate the collection of retail transaction data, and membership in a customer loyalty program can further associate the transaction with a specific customer. In this section, we discuss the development of probabilistic if–then statements, called **association rules**, which convey the likelihood of certain items being purchased together. Although association rules are an important tool in **market basket analysis**, they are also applicable to disciplines other than marketing. For example, association rules can assist medical researchers in understanding which treatments have been commonly prescribed to certain patient symptoms (and the resulting effects of those treatments).

Hy-Vee grocery store would like to gain insight into its customers' purchase patterns to possibly improve its in-aisle product placement and cross-product promotions. Table 6.6 contains a small sample of data in which each transaction comprises the items purchased by a shopper in a single visit to a Hy-Vee. An example of an association rule from this data would be "if {bread, jelly}, then {peanut butter}," meaning that "if a transaction includes bread and jelly, then it also includes peanut butter." The collection of items (or **itemset**) corresponding to the *if* portion of the rule, {bread, jelly}, is called the **antecedent**. The itemset corresponding to the *then* portion of the rule, {peanut butter}, is called the **consequent**.

Typically, only association rules for which the consequent consists of a single item are considered because these are more actionable. Although the number of possible association rules can be overwhelming, we typically investigate only association rules that involve antecedent and consequent itemsets that occur together frequently. To formalize the notion of "frequent," we define the **support** of an itemset as the percentage of transactions in the data that include that itemset. In Table 6.6, the support of {bread, jelly} is 4/10 = 0.4. For a transaction randomly selected from the data set displayed in Table 6.6, the probability of it containing the itemset {bread, jelly} is 0.4. The potential impact of an association rule is often governed by the number of transactions it may affect, which is measured by computing the support of the itemset consisting of the union of its antecedent and consequent. Investigating the rule "if {bread, jelly}, then {peanut butter}" from Table 6.6, we see the support of {bread, jelly, peanut butter} is 2/10 = 0.2. For a transaction randomly selected from the data set displayed in Table 6.6, the probability of it containing the itemset {bread, jelly, peanut butter} is 0.2. By only considering rules involving itemsets with a support above a minimum level, inexplicable rules capturing random noise in the data can generally be avoided. A rule of thumb

Support is also sometimes expressed as the number (or count) of total transactions in the data containing an itemset.

The data in Table 6.6 are in item-list format; that is, each transaction row corresponds to a list of item names. Alternatively, the data can be represented in binary matrix format, in which each row is a transaction record and the columns correspond to each distinct item. A third approach is to store the data in stacked form in which each row is an ordered pair; the first entry is the transaction number and the second entry is the item.

Table 6.6 Shopping-Cart Transactions

Transaction	Shopping Cart
1	bread, peanut butter, milk, fruit, jelly
2	bread, jelly, soda, potato chips, milk, fruit, vegetables, peanut butter
3	whipped cream, fruit, chocolate sauce, beer
4	steak, jelly, soda, potato chips, bread, fruit
5	jelly, soda, peanut butter, milk, fruit
6	jelly, soda, potato chips, milk, bread, fruit
7	fruit, soda, potato chips, milk
8	fruit, soda, peanut butter, milk
9	fruit, cheese, yogurt
10	yogurt, vegetables, beer

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

is to consider only association rules with a support of at least 20% of the total number of transactions. If an itemset is particularly valuable and represents a lucrative opportunity, then the minimum support used to filter the rules can be lowered.

A property of a reliable association rule is that, given a transaction contains the antecedent itemset, there is a high probability that it contains the consequent itemset. This conditional probability of *P*(consequent itemset | antecedent itemset) is called the **confidence** of a rule, and is computed as

The definition of confidence follows from the definition of conditional probability discussed in Chapter 4.

Confidence

$$P(\text{consequent} \mid \text{antecedent}) = \frac{P(\text{consequent and antecedent})}{P(\text{antecedent})} = \frac{\text{support of}\{\text{consequent and antecedent}\}}{\text{support of antecedent}}$$

Although high value of confidence suggests a rule in which the consequent is frequently true when the antecedent is true, a high value of confidence can be misleading. For example, if the support of the consequent is high—that is, the itemset corresponding to the *then* part is very frequent—then the confidence of the association rule could be high even if there is little or no association between the items. In Table 6.6, the rule "if {cheese}, then {fruit}" has a confidence of 1.0 (or 100%). This is misleading because {fruit} is a frequent item; *almost any* rule with {fruit} as the consequent will have high confidence. Therefore, to evaluate the efficiency of a rule, we need to compare the *P*(consequent | antecedent) to the *P*(consequent) to determine how much more likely the consequent itemset is given the antecedent itemset versus just the overall (unconditional) likelihood that a transaction contains the consequent. The ratio of *P*(consequent | antecedent) to *P*(consequent) is called the **lift ratio** of the rule and is computed as:

Lift Ratio

$$\frac{P(\text{consequent} \mid \text{antecedent})}{P(\text{consequent})} = \frac{P(\text{consequent and antecedent})}{P(\text{consequent}) \times P(\text{antecedent})} = \frac{\text{confidence of rule}}{\text{support of consequent}}$$

Thus, the lift ratio represents how effective an association rule is at identifying transactions in which the consequent itemset occurs versus a randomly selected transaction. A lift ratio greater than one suggests that there is some usefulness to the rule and that it is better at identifying cases when the consequent occurs than having no rule at all. From the definition of lift ratio, we see that the denominator contains the probability of a transaction containing the consequent set multiplied by the probability of a transaction containing the antecedent set. This product of probabilities is equivalent to the expected likelihood of a transaction containing both the consequent itemset and the antecedent itemset if these itemsets were independent. In other words, a lift ratio greater than one suggests that the level of association between the antecedent and consequent is higher than would be expected if these itemsets were independent.

For the data in Table 6.6, the rule "if {bread, jelly}, then {peanut butter}" has confidence = 2/4 = 0.5 and lift ratio = 0.5/0.4 = 1.25. In other words, a customer who purchased both bread and jelly is 25% more likely to have purchased peanut butter than a randomly selected customer.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The utility of a rule depends on both its support and its lift ratio. Although a high lift ratio suggests that the rule is very efficient at finding when the consequent occurs, if it has a very low support, the rule may not be as useful as another rule that has a lower lift ratio but affects a large number of transactions (as demonstrated by a high support). However, an association rule with a high lift ratio and low support may still be useful if the consequent represents a very valuable opportunity.

DATA *file*
hyvee_binary.csv
hyvee_stacked.csv

Based on the data in Table 6.6, Table 6.7 shows the list of association rules that achieve a lift ratio of at least 1.39 while satisfying a minimum support of 40% and a minimum confidence of 50%. The top rules in Table 6.7 suggest that bread, fruit, and jelly are commonly associated items. For example, the fourth rule listed in Table 6.7 states, "If Fruit and Jelly are purchased, then Bread is also purchased." Perhaps Hy-Vee could consider a promotion and/or product placement to leverage this perceived relationship.

Evaluating Association Rules

For an association rule to be useful, it must occur often enough to be valuable as a business opportunity (large enough support), it must be likely to be true (high enough confidence), and it must be surprising and not due to just chance (large lift ratio). In particular, association rules with a weak support measure often have no logical explanation that can yield insights for decision making. The support of an association rule can generally be improved

Table 6.7 Association Rules for Hy-Vee

Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A & C	Confidence	Lift Ratio
Bread	Fruit, Jelly	0.40	0.50	0.40	1.00	2.00
Bread	Jelly	0.40	0.50	0.40	1.00	2.00
Bread, Fruit	Jelly	0.40	0.50	0.40	1.00	2.00
Fruit, Jelly	Bread	0.50	0.40	0.40	0.80	2.00
Jelly	Bread	0.50	0.40	0.40	0.80	2.00
Jelly	Bread, Fruit	0.50	0.40	0.40	0.80	2.00
Fruit, Potato Chips	Soda	0.40	0.60	0.40	1.00	1.67
Peanut Butter	Milk	0.40	0.40	0.60	1.00	1.67
Peanut Butter	Milk, Fruit	0.40	0.60	0.40	1.00	1.67
Peanut Butter, Fruit	Milk	0.40	0.60	0.40	1.00	1.67
Potato Chips	Fruit, Soda	0.40	0.60	0.40	1.00	1.67
Potato Chips	Soda	0.40	0.60	0.40	1.00	1.67
Fruit, Soda	Potato Chips	0.60	0.40	0.40	0.67	1.67
Milk	Peanut Butter	0.60	0.40	0.40	0.67	1.67
Milk	Peanut Butter, Fruit	0.60	0.40	0.40	0.67	1.67
Milk, Fruit	Peanut Butter	0.60	0.40	0.40	0.67	1.67
Soda	Fruit, Potato Chips	0.60	0.40	0.40	0.67	1.67
Soda	Potato Chips	0.60	0.40	0.40	0.67	1.67
Fruit, Soda	Milk	0.60	0.60	0.50	0.83	1.39
Milk	Fruit, Soda	0.60	0.60	0.50	0.83	1.39
Milk	Soda	0.60	0.60	0.50	0.83	1.39
Milk, Fruit	Soda	0.60	0.60	0.50	0.83	1.39
Soda	Milk	0.60	0.60	0.50	0.83	1.39
Soda	Milk, Fruit	0.60	0.60	0.50	0.83	1.39

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

by basing it on less specific antecedent and consequent itemsets. For example, the support of association rules based on granular items *sourdough bread, whole wheat bread, white bread*, etc. will be smaller than association rules based on the aggregate item *bread*. However, the association rules based on the less specific item (bread) may yield less insight than the association rules based on the more granular items (*sourdough bread, whole wheat bread, white bread*). Adjusting the data by aggregating items into more general categories (or splitting items into more specific categories) so that items occur in roughly the same number of transactions often yields better association rules.

Although explicit measures such as support, confidence, and lift ratio can help filter association rules, an association rule is ultimately judged on how actionable it is and how well it explains a previously unknown relationship between itemsets. For example, suppose the retail chain Walmart mined its transactional data to uncover strong evidence of the association rule, "If a customer purchases a protein bar, then a customer also purchases an energy drink." Walmart could leverage this relationship in product placement decisions as well as in advertisements and promotions, perhaps by placing a high-margin energy drink display near the protein bars. However, we must be aware that association rule analysis often results in obvious relationships such as "If a customer purchases hamburger patties, then a customer also purchases hamburger buns," which may be true but provide no new insight. It may be possible to improve the utility of association rules by introducing a **virtual item** into the transaction data. A virtual item is not explicitly part of the transaction. It is not a product purchased by a customer, but represents a feature that may be important in associating the item to other purchased items. For example, in a purchase transaction, possible virtual items may be the time of the transaction, location of the transaction, sales representative involved in transaction, and any known information about the customer. For example, Walmart could include the relative time during the day of the transaction (morning, afternoon, evening) as a virtual item in its analysis of protein bar and energy drink purchases. This inclusion of the virtual item could reveal that purchases of a protein bar in the morning leads to purchases of an energy drink with much more frequency than purchases of a protein bar in the evening.

6.4 Text Mining

Every day, nearly 500 million tweets are published using the online social network service Twitter. Many of these tweets contain important clues about how Twitter users value a company's products and services. Some tweets might sing the praises of a product; others might complain about low-quality service. Furthermore, Twitter users vary greatly in the number of followers (some have millions of followers and others just a few) and therefore these users have varying degrees of influence. Data-savvy companies can use social media data to improve their products and services. Online reviews on web sites such as Amazon and Yelp also provide data on how customers feel about products and services.

However, the data in these examples are not numerical. The data are text: words, phrases, sentences, and paragraphs. Text, like numerical data, may contain information that can help solve problems and lead to better decisions. **Text mining** is the process of extracting useful information from text data. In this section, we discuss text mining, how it is different from data mining of numerical data, and how it can be useful for decision making.

Text data is often referred to as **unstructured data** because in its raw form, it cannot be stored in a traditional structured database (with observations in rows and variables in columns). Audio and video data are also examples of unstructured data. Data mining with text data is more challenging than data mining with traditional numerical data, because it requires more preprocessing to convert the text to a format amenable for analysis. However, once the text data has been converted to numerical data, we can apply the data mining methods discussed earlier in this chapter. We begin with a small example which illustrates how text data can be converted to numerical data and then analyzed. Then we will provide more in-depth discussion of text-mining concepts and preprocessing procedures.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Voice of the Customer at Triad Airlines

Triad Airlines is a regional commuter airline. Through its "voice of the customer" program, Triad solicits feedback from its customers through a follow-up e-mail the day after the customer has completed a flight. The e-mail survey asks the customer to rate various aspects of the flight and asks the respondent to type comments into a dialog box in the e-mail.

In addition to the quantitative feedback from the ratings, the comments entered by the respondents need to be analyzed so that Triad can better understand its customers' specific concerns and respond in an appropriate manner. Table 6.8 contains a small sample of these comments we will use to illustrate how descriptive data mining can be used in this business context. In the text mining domain, a contiguous piece of text is referred to as a **document**. A document can be a single sentence or an entire book, depending on how the text is organized for analysis. Each document is composed of individual **terms**, which often correspond to words. In general, a collection of text documents to be analyzed is called a **corpus**. In the Triad Airlines example, our corpus consists of 10 documents, where each document is a single customer's comments.

Triad's management would like to categorize these customer comments into groups whose member comments share similar characteristics so that a focused solution team can be assigned to each group of comments.

Preprocessing text can be viewed as determining how to represent the features for a subsequent descriptive or predictive data mining method. To be analyzed, text data needs to be converted to structured data (rows and columns of numerical data) so that the tools of descriptive statistics, data visualization, and data mining can be applied. Considering each document as an observation (row in a data set), we wish to represent the text in the document with variables (or columns in a data set). One common approach, called **bag of words**, treats every document as just a collection of individual words (or terms). Bag of words is a simple, but often effective, approach that ignores natural language processing aspects such as grammar, word order, and sentence structure. In bag of words, we can think of converting a group of documents into a matrix of rows and columns where the rows correspond to a document and the columns correspond to a particular word. In Triad's case, a document is a single respondent's comment. A **presence/absence document-term matrix** is a matrix with the rows representing documents and the columns representing words, and the entries in the columns indicating either the presence or the absence of a particular word in a particular document (1 = present and 0 = not present). This matrix is also called **binary document-term matrix**.

If the document-term matrix is transposed (so that the terms are in the rows and the documents are in the columns), the resulting matrix is referred to as a term-document matrix

Creating the list of terms to use in the presence/absence matrix can be a complicated matter. Too many terms results in a matrix with many columns, which may be difficult to manage and could yield meaningless results. Too few terms may miss important relationships.

Table 6.8 Ten Respondents' Comments for Triad Airlines

Comments
The wi-fi service was horrible. It was slow and cut off several times.
My seat was uncomfortable.
My flight was delayed 2 hours for no apparent reason.
My seat would not recline.
The man at the ticket counter was rude. Service was horrible.
The flight attendant was rude. Service was bad.
My flight was delayed with no explanation.
My drink spilled when the guy in front of me reclined his seat.
My flight was canceled.
The arm rest of my seat was nasty.

DATA*file*
triad.csv

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Often, term frequency along with the problem context are used as a guide in identifying the terms to include in a document-term matrix. We discuss this in more detail in the next section. In Triad's case, management used word frequency and their customer service experience to devise the following list of terms they feel are relevant for categorizing the respondent's comments: delayed, flight, horrible, recline, rude, seat, and service.

As shown in Table 6.9, these 7 terms correspond to the columns of the presence/absence document-term matrix and the rows correspond to the 10 documents. Each matrix entry indicates whether or not a column's term appears in the document corresponding to the row. For example, a 1 entry in the first row and third column means that the term "horrible" appears in document 1. A zero entry in the third row and fourth column means that the term "recline" does not appear in document 3.

Having converted the text to numerical data, we can apply clustering. In this case, because we have binary presence-absence data, we apply hierarchical clustering. Observing that the absence of a term in two different documents does not imply similarity between the documents, we select Jaccard's distance to measure dissimilarity between observations (documents). To measure dissimilarity between clusters, we use the complete linkage agglomeration method. At the level of three clusters, hierarchical clustering results in the following groups of documents:

Cluster 1: {1, 5, 6} = documents discussing service issues
Cluster 2: {2, 4, 8, 10} = documents discussing seat issues
Cluster 3: {3, 7, 9} = documents discussing schedule issues

With these three clusters defined, management can assign an expert team to each of these clusters to directly address the concerns of its customers.

Preprocessing Text Data for Analysis

In general, the text mining process converts unstructured text into numerical data and applies data mining techniques. For the Triad example, we converted the text documents into a binary document-term matrix and then applied hierarchical clustering to gain insight on the different types of comments. In this section, we present a more detailed discussion of terminology and methods used in preprocessing text data into numerical data for analysis.

Converting documents to a document-term matrix is not a simple task. Obviously, which terms become the headers of the columns of the document-term matrix can greatly impact the analysis. **Tokenization** is the process of dividing text into separate terms, referred to

Table 6.9 The Presence/Absence Document-Term Matrix for Triad Airlines

	Term						
Document	Delayed	Flight	Horrible	Recline	Rude	Seat	Service
1	0	0	1	0	0	0	1
2	0	0	0	0	0	1	0
3	1	1	0	0	0	0	0
4	0	0	0	1	0	1	0
5	0	0	1	0	1	0	1
6	0	1	0	0	1	0	1
7	1	1	0	0	0	0	0
8	0	0	0	1	0	1	0
9	0	1	0	0	0	0	0
10	0	0	0	0	0	1	0

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

as tokens. The process of identifying tokens is not straightforward and involves **term normalization**, a set of natural language processing techniques to map text into a standardized form. First, symbols and punctuations must be removed from the document and all letters should be converted to lowercase. For example, "Awesome!", "awesome," and "#Awesome" should all be converted to "awesome." Likewise, different forms of the same word, such as "stacking," "stacked," and "stack" probably should not be considered as distinct terms. **Stemming**, the process of converting a word to its stem or root word, would drop the "ing" and "ed" suffixes and place only "stack" in the list of terms to be tracked.

The goal of preprocessing is to generate a list of most relevant terms that is sufficiently small so as to lend itself to analysis. In addition to stemming, frequency can be used to eliminate words from consideration as tokens. For example, if a term occurs very frequently in every document in the corpus, then it probably will not be very useful and can be eliminated from consideration. Text mining software contains procedures to automatically remove very common words, called **stopwords**, such as "the," "and," and "of." Similarly, low-frequency words probably will not be very useful as tokens. Another technique for reducing the consideration set for tokens is to consolidate a set of words that are synonyms. For example, "courteous," "cordial," and "polite" might be best represented as a single token, "polite."

In addition to automated stemming and text reduction via frequency and synonyms, most text-mining software gives the user the ability to manually specify terms to include or exclude as tokens. Also, the use of slang, humor, irony, and sarcasm can cause interpretation problems and might require more sophisticated data cleansing and subjective intervention on the part of the analyst to avoid misinterpretation.

Data preprocessing parses the original text data down to the set of tokens deemed relevant for the topic being studied. Based on these tokens, a presence/absence document-term matrix such as in Table 6.9 can be generated.

When the documents in a corpus contain many more words than the brief comments in the Triad Airline example, and when the *frequency* of word occurrence is important to the context of the business problem, preprocessing can be used to develop a **frequency document-term matrix**. A frequency document-term matrix is a matrix whose rows represent documents and columns represent tokens, and the entries in the matrix are the frequency of occurrence of each token in each document. We illustrate this in the following example.

Movie Reviews

A new action film has been released and we now have a sample of 10 reviews from movie critics. Using preprocessing techniques, including text reduction by synonyms, we have reduced the number of tokens to only two: "great" and "terrible." Table 6.10 displays the corresponding frequency document-term matrix. As Table 6.10 shows, the token "great" appears four times in Document 7. Reviewing the entire table, we observe that five is the maximum frequency of a token in a document and zero is the minimum frequency.

To demonstrate the analysis of a frequency document-term matrix with descriptive data mining, we apply k-means clustering with $k = 2$ to the frequency document-term matrix to obtain the two clusters in Figure 6.21. Cluster 1 contains reviews that tend to be negative and Cluster 2 contains reviews that tend to be positive. We note that the Observation (3, 3) corresponds to the balanced review of Document 4; based on this small corpus, the balanced review is more similar to the positive reviews than the negative reviews, suggesting that the negative reviews may tend to be more extreme.

Table 6.10 shows the raw counts (frequencies) of terms. When documents in a corpus substantially vary in length (number of terms), it is common to adjust for document length by dividing the raw term frequencies by the total number of terms in the document.

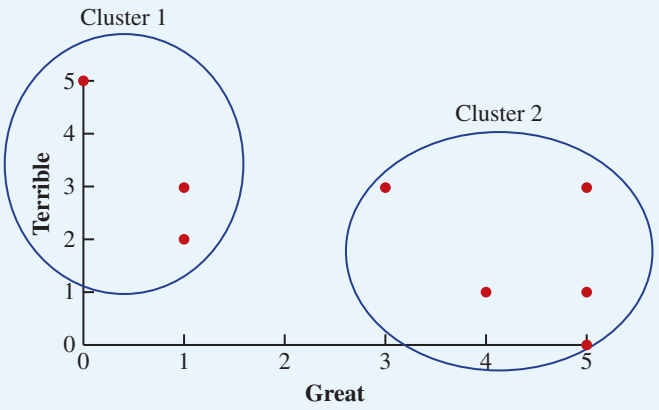
Term frequency (whether based on raw count or relative frequency to account for document length) is a text mining measure that pertains to an individual document. For a particular document, an analyst may also be interested in how unique a term is relative to the other documents in the corpus. As previously mentioned, it is common to impose lower and

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 6.10 The Frequency Document-Term Matrix for Movie Reviews

	Term	
Document	Great	Terrible
1	5	0
2	5	1
3	5	1
4	3	3
5	5	1
6	0	5
7	4	1
8	5	3
9	1	3
10	1	2

Figure 6.21 Grouping Movie Reviews Using *k*-Means Clustering with $k = 2$



upper limits on the term frequencies to filter out terms that are extremely rare or extremely common. In addition to those mechanisms, an analyst may be interested in weighting terms based on their distribution over the corpus. The logic is that the fewer the documents in which a term occurs, the more likely it is to be potentially insightful to the analysis of the documents it does occur in.

A popular measure for weighting terms based on frequency and uniqueness is **term frequency times inverse document frequency (TFIDF)**. The TFIDF of a term t in document d is computed as follows:

$$\begin{aligned} TFIDF(t,d) &= \text{term frequency} \times \text{inverse document frequency} \\ &= (\text{number of times term } t \text{ appears in document } d) \\ &\quad \times \left[1 + \ln\left(\frac{\text{total number of documents in corpus}}{\text{number of documents containing term } t}\right)\right] \end{aligned}$$

The inverse document frequency portion of the TFIDF calculation gives a boost to terms for being unique. Thus, using $TFIDF(t, d)$ more highly scores a term t that

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

frequently appears in document d and does not appear frequently in other documents. Thus, basing a document-term matrix on TFIDF can make the unique terms (with respect to their frequency) more pronounced.

For five different possible term frequencies of term t in a document d (represented by $TF(t,d)$), Figure 6.22 displays the $TFIDF(t,d)$ value as the number of total documents containing the term t ranges from one document to 100 documents (in a corpus of 100 documents).

To better understand how TFIDF weights terms, first consider five terms t_1, t_2, t_3, t_4, and t_5, that appear 1, 2, 3, 4, and 5 times, respectively, in a document d. Then the TFIDF for each of these terms with respect to document d corresponds to one of the five curves in Figure 6.22. Suppose each of these five terms also appear in all 99 of the other documents. Then,

$$TFIDF\,(t_1,d) = 1 \times \left[1 + \ln\left(\frac{100}{100}\right)\right] = 1$$

$$TFIDF\,(t_2,d) = 2 \times \left[1 + \ln\left(\frac{100}{100}\right)\right] = 2$$

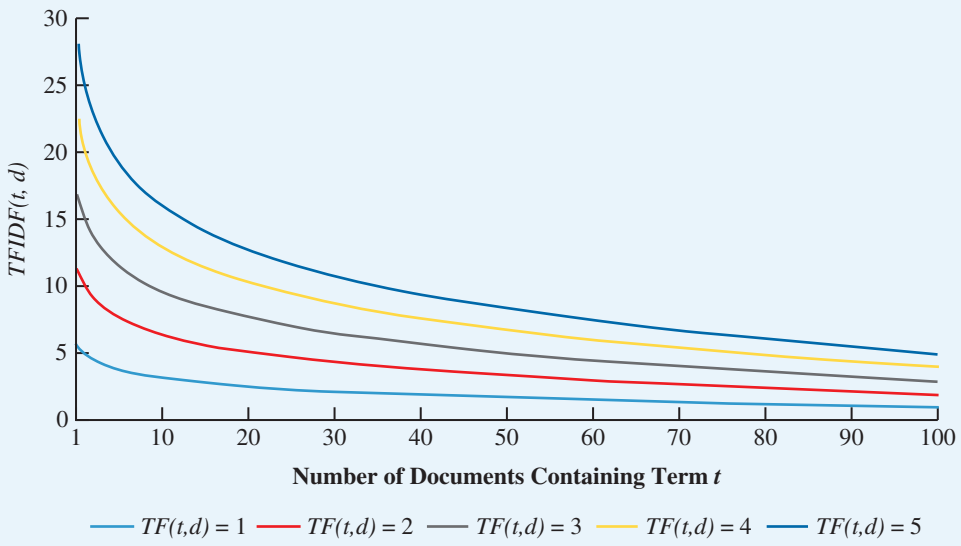
$$TFIDF\,(t_3,d) = 3 \times \left[1 + \ln\left(\frac{100}{100}\right)\right] = 3$$

$$TFIDF\,(t_4,d) = 4 \times \left[1 + \ln\left(\frac{100}{100}\right)\right] = 4$$

$$TFIDF\,(t_5,d) = 5 \times \left[1 + \ln\left(\frac{100}{100}\right)\right] = 5$$

These five values of $TFIDF(t,d)$ correspond to the farthest right values on the five curves in Figure 6.22.

Figure 6.22 *TFIDF(t,d)* for Varying Levels of Term Frequency and Corpus Uniqueness



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To see how TFIDF weights uniqueness of a term in a corpus. consider two terms t_6 and t_7 that each occur once in a document d. However, term t_6 does not occur in any of the other 99 documents and t_7 occurs in all the other 99 documents. Then,

$$TFIDF\,(t_6,d) = 1 \times \left[1 + \ln\left(\frac{100}{1}\right)\right] = 5.61$$

$$TFIDF\,(t_7,d) = 1 \times \left[1 + \ln\left(\frac{100}{100}\right)\right] = 1.00$$

These two values of *TFIDF(t,d)* correspond to the farthest left value and farthest right value on the *TF(t, d)* = 1 curve in Figure 6.22. Based on term frequency, terms t_6 and t_7 are weighted the same (with a value of 1), but according to TFIDF, the rarer occurrence of term t_6 in other documents gives it over five times the weight of term t_7.

Now consider two terms t_8 and t_9 that each occur five times in a document d. However, term t_8 does not occur in any of the other 99 documents and t_9 occurs in all the other 99 documents. Then,

$$TFIDF\,(t_8,d) = 5 \times \left[1 + \ln\left(\frac{100}{1}\right)\right] = 28.03$$

$$TFIDF\,(t_9,d) = 5 \times \left[1 + \ln\left(\frac{100}{100}\right)\right] = 5.00$$

We see again that while according to term frequency, terms t_8 and t_9 are weighted the same (5), but according to TFIDF, the rarer occurrence of term t_8 in other documents gives it over five times the weight of term t_9.

To see how TFIDF accounts for both term frequency and corpus uniqueness, compare $TFIDF\,(t_8,d) = 28.03$ to $TFIDF\,(t_7,d) = 1$. Term t_8, which is unique to document d and occurs a relatively frequent five times in document d, has a weight over 28 times the weight of term t_7, which occurs in all 100 documents of the corpus and occurs only once in document d.

Computing Dissimilarity Between Documents

After preprocessing text and the conversion of a corpus into a frequency document-term matrix (based on raw frequency, relative frequency, or TFIDF), the notion of distance between observations discussed earlier in this chapter applies. In this case, the distance between observations corresponds to the dissimilarity between documents. To measure the dissimilarity between text documents, **cosine distance** is commonly used. Cosine distance is computed by:

$$d_{uv}^{\cos} = 1 - \frac{u_1 \times v_1 + \cdots + u_q \times v_q}{\sqrt{u_1^2 + \cdots u_q^2}\sqrt{v_1^2 + \cdots v_q^2}}$$

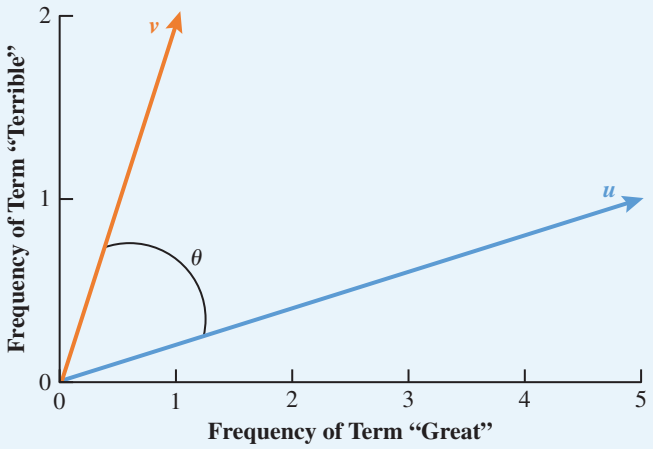
We demonstrate the calculation of cosine distance by considering document 3 and document 10 in Table 6.10. Referring to Table 6.10, document 3 contains the term "great" five times and the term "terrible" once. Therefore, $u_1 = 5$ and $u_2 = 1$. Document 10 contains the term "great" once and the term "terrible" twice, so Therefore, $v_1 = 1$ and $v_2 = 2$. Then, the cosine distance between document 3 and document 10 is:

$$d_{uv}^{\cos} = 1 - \frac{5 \times 1 + 1 \times 2}{\sqrt{5^2 + 1^2}\sqrt{1^2 + 2^2}} = 0.386$$

To visualize the cosine distance between two observations, in this case $u = (5, 1)$ and $v = (1, 2)$, Figure 6.23 represents these observations as vectors emanating from the origin.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6.23 Visualization of Cosine Distance



The cosine distance between two observations is equivalent to the cosine of the angle (measured in radians) between their corresponding vectors. Cosine distance can be particularly useful for analyzing a frequency document-term matrix because the angle between two observation vectors does not depend on the magnitude of the variables (making it different than distance measures discussed earlier in this chapter). This allows cosine distance to measure dissimilarity in frequency patterns rather than frequency magnitude.

To see how magnitude does not affect cosine distance, consider a new movie review document that contains the term "great" 10 times and the term "terrible" twice. This new document can be represented by the vector $w = (10, 2)$. The cosine distance between document 3 ($u = (5, 1)$) and the new movie review document ($w = (10, 2)$) is

$$d_{uw}^{\cos} = 1 - \frac{5 \times 10 + 1 \times 2}{\sqrt{5^2 + 1^2}\sqrt{10^2 + 2^2}} = 0$$

Notice that although the new movie document, $w = (10, 2)$, has twice as many mentions of "great" and twice as many mentions of "terrible" as document 3, $u = (5, 1)$, these two documents are considered identical because the relative mentions of "great" to mentions of "terrible" are the same, $10/2 = 5/1 = 5$. Figure 6.24 illustrates that the angle between $u = (5, 1)$ and $w = (10, 2)$ is zero and from trigonometry, we know $\cos 0 = 0$.

Furthermore, the cosine distance between the new movie review document ($w = (10, 2)$) and document 10 ($v = (1, 2)$) is

$$d_{wv}^{\cos} = 1 - \frac{10 \times 1 + 2 \times 2}{\sqrt{10^2 + 2^2}\sqrt{1^2 + 2^2}} = 0.386$$

Thus, the cosine distance between $w = (10, 2)$ and $v = (1, 2)$ is the same as the cosine distance between $u = (5, 1)$ and $v = (1, 2)$. Figure 6.24 shows this equivalence by showing that the angle between v and u is the same as the angle between v and w.

Word Clouds

A **word cloud** is a visual representation of a document or set of documents in which the size of the word is proportional to the frequency with which the word appears. Figure 6.25 displays a word cloud of Chapter 1 of this textbook.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 6.24 Cosine Distance Measuring Dissimilarity in Pattern Not Magnitude

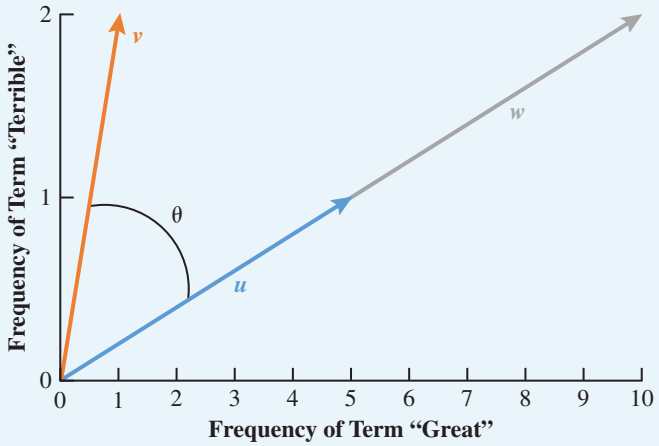
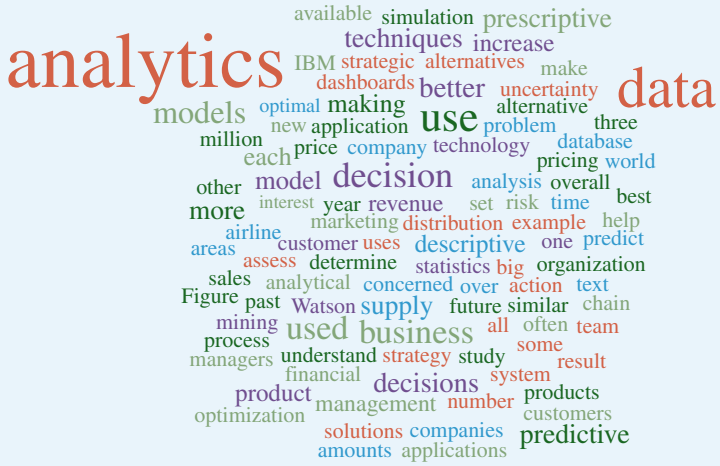


Figure 6.25 A Word Cloud of Chapter 1 of this Text



As can be seen from this word cloud, analytics and data are used most frequently in Chapter 1. Other more frequently mentioned words are decision, models, prescriptive, and predictive. The word cloud gives a quick visual sense of what the document's content. Using word clouds on tweets, for example, can provide insight on trending topics.

Summary

We have introduced descriptive data-mining methods and related concepts. We began the chapter with a discussion of dimension reduction. We illustrated the method of principal component analysis and showed how it can be used to potentially reduce the dimensionality of a dataset. We then introduced clustering analysis as a form of descriptive data mining. After introducing how to measure distance between individual observations, we

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

presented two methods for grouping observations based on the similarity of their respective variable values: *k*-means clustering and hierarchical clustering. In *k*-means clustering, the analyst specifies the number of clusters (k) and then places each observation into one of the k clusters in an attempt to minimize the dissimilarity within the clusters. Agglomerative hierarchical clustering begins with each observation in its own cluster and iteratively aggregates clusters using a specified agglomeration method. We described several of these agglomeration methods and discussed their characteristics. We concluded our discussion of clustering with a comparison of *k*-means clustering and hierarchical clustering.

We then introduced association rules and explained their use for identifying patterns across transactions, particularly in retail data. We defined the concepts of support count, confidence, and lift ratio, and we described their utility in gleaning actionable insight from association rules.

Finally, we discussed the text-mining process. Specifically, we discuss the bag of words approach which treats individual words as features. Text is first preprocessed by deriving a smaller set of tokens from the larger set of words contained in a collection of documents. The tokenized text data is then converted into a presence/absence document-term matrix or a frequency document-term matrix. We then discuss term frequency times inverse document frequency as an alternative way to weight terms and the use of cosine distance to measure the dissimilarity between two documents based on differences in the relative patterns of term usage. We then demonstrated the application of hierarchical clustering on a binary document-term matrix and *k*-means clustering on a frequency document-term matrix to glean insight from the underlying text data.

Glossary

Antecedent The itemset corresponding to the *if* portion of an if–then association rule.
Association rule An if–then statement describing the relationship between itemsets.
Bag of words An approach for processing text into a structured row-column data format in which documents correspond to row observations and words (or more specifically, terms) correspond to column variables.
Binary document-term matrix A matrix with the rows representing documents (units of text) and the columns representing terms (words or word roots), and the entries in the columns indicating either the presence or absence of a particular term in a particular document (1 = present and 0 = not present).
Cluster cohesion A concept describing the distance between observations in the same cluster.
Cluster interpretability A concept describing how well a cluster can be explained within the context of its application.
Cluster separation A concept describing how distinct a cluster is from other clusters.
Cluster stability A concept describing how robust a set of clusters is to slight changes in the data.
Complete linkage A measure of calculating dissimilarity between clusters by considering only the two most dissimilar observations between the two clusters.
Confidence The conditional probability that the consequent of an association rule occurs given the antecedent occurs.
Consequent The itemset corresponding to the *then* portion of an if–then association rule.
Corpus A collection of documents to be analyzed.
Cosine distance A measure of dissimilarity between two observations often used on frequency data derived from text because it is unaffected by the magnitude of the frequency and instead measures differences in frequency patterns.
Dendrogram A tree diagram used to illustrate the sequence of nested clusters produced by hierarchical clustering.
Dimension reduction The process of reducing the number of variables in a data set by constructing new variables which are weighted combinations of the original variables.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Document A piece of text, which can range from a single sentence to an entire book depending on the scope of the corresponding corpus.
Elbow method An approach to assess sets of clusters by charting a measure of cluster cohesion for varying numbers of clusters and identifying the point at which the marginal change in cluster cohesion diminishes.
Euclidean distance A geometric measure of dissimilarity between observations based on the Pythagorean theorem.
Feature (variable) selection The process of reducing the number of variables to consider for further analysis by selecting a subset of the original variables.
Frequency document-term matrix A matrix whose rows represent documents (units of text) and columns represent terms (words or word roots), and the entries in the matrix are the number of times each term occurs in each document.
Group average linkage A measure of calculating dissimilarity between clusters by considering the distance between each pair of observations between two clusters.
Hierarchical clustering Process of agglomerating observations into a series of nested groups based on a measure of similarity.
Itemset A collection of items in the composing the antecedent and/or consequent of an association rule.
Jaccard's coefficient A measure of similarity between observations consisting solely of binary categorical variables that considers only matches of nonzero entries.
Jaccard distance A measure of dissimilarity between observations based on Jaccard's coefficient.
***k*-means clustering** Process of organizing observations into one of k groups based on a measure of similarity (typically Euclidean distance).
Lift ratio The ratio of the performance of a data mining model measured against the performance of a random choice. In the context of association rules, the lift ratio is the ratio of the probability of the consequent occurring in a transaction that satisfies the antecedent versus the probability that the consequent occurs in a randomly selected transaction.
Manhattan distance A measure of dissimilarity between two observations based on the sum of the absolute differences in each variable dimensions.
Market basket analysis Analysis of items frequently co-occurring in transactions (such as purchases).
Market segmentation The partitioning of customers into groups that share common characteristics so that a business may target customers within a group with a tailored marketing strategy.
Matching coefficient A measure of similarity between observations based on the number of matching values of categorical variables.
Matching distance A measure of dissimilarity between observations based on the matching coefficient.
Observation (record) A set of observed values of variables associated with a single entity, often displayed as a row in a spreadsheet or database.
Presence/absence document-term matrix A matrix with the rows representing documents and the columns representing words, and the entries in the columns indicating either the presence or the absence of a particular word in a particular document (1 = present and 0 = not present).
Principal component analysis A dimension-reduction method that sequentially constructs a set of variables which are linear combinations of the original variables but are independent of each other.
Scree plot A line chart that displays the proportion of variance explained by each principal component.
Silhouette A measure of comparing how similar an observation is to other observations in its own cluster versus observations in other clusters.
Single linkage A measure of calculating dissimilarity between clusters by considering only the two most similar observations between the two clusters.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Stemming The process of converting a word to its stem or root word.
Stopwords Common words in a language that are removed in the preprocessing of text.
Support The percentage of transactions in which a collection of items occurs together in a transaction data set.
Term The most basic unit of text comprising a document, typically corresponding to a word or word stem.
Term frequency times inverse document frequency (TFIDF) Text mining measure which accounts for term frequency and the uniqueness of a term in a document relative to other documents in a corpus.
Term normalization A set of natural language processing techniques to map text into a standardized form.
Text mining The process of extracting useful information from text data.
Tokenization The process of dividing text into separate terms, referred to as tokens.
Unsupervised learning A category of data-mining techniques in which an algorithm explains relationships without an outcome variable to guide the process.
Unstructured data Data, such as text, audio, or video, that cannot be stored in a traditional structured database.
Virtual item An item introduced to transactional data that is not explicitly part of the transaction, but represents a feature that may be provide insight as part of an association rule
Ward's linkage A procedure that aggregates observations in a manner to obtain clusters with the least amount of information loss due to the aggregation.
Word cloud A visualization of text data based on word frequencies in a document or set of documents.

Problems

Problems 1 through 26 and Case Problem 1 do not require the use of data mining software and focus on knowledge of concepts and basic calculations.

Conceptual

1. **Properties of Principal Component Analysis.** Which of the following statements about principal component analysis is true? LO 1
   a. A strength of PCA is the interpretability of its output.
   b. At least four components are necessary to satisfactorily explain the variance in a data set.
   c. PCA is a predictive method.
   d. Each additional principal component generated is independent of the proceeding principal components, but describes a decreasing amount of the variance in the input variables.
2. **Distance Measure for Binary Variables.** What is an appropriate choice of distance metric for observations consisting of binary variables for which matching zeroes do not correspond to similarity? LO 3
   a. Matching distance
   a. Jaccard distance
   b. Euclidean distance
   c. Cosine distance
3. **Cosine Distance Application.** When may cosine distance be a good choice to measure differences between observations? LO 2, 3
   a. When dealing with observations with variables with both quantitative and categorical variables.
   b. When dealing with observations consisting of binary variables.
   c. When similarity in the pattern of variable values of two observations is more relevant to the application than the similarity in terms of the magnitude of variable values.
   d. When dealing with observations consisting of ordinal variables.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4. **Euclidean versus Manhattan Distance.** Which of the following is a true statement about the comparison of Euclidean distance versus Manhattan distance? LO 2
   a. Manhattan distance is distorted less by outlier observations.
   b. Euclidean distance is more applicable to binary variables.
   c. Manhattan distance scales better to higher dimensions.
   d. Euclidean distance is less expensive to compute.
5. **Properties of Hierarchical Clustering.** Which of the following statements about hierarchical clustering is true? LO 5, 6
   a. Because of its nested design, hierarchical clustering will find the set of k clusters that maximizes the average silhouette meaure for all possible values of k.
   b. For the same number of clusters and same distance measure, the clusters resulting from a k-means approach and a hierarchical approach will be the same.
   c. Hierarchical clustering is not affected by outliers.
   d. A dendrogram shows how to group observations into varying numbers of clusters.
6. **Assessing *k*-Means Clusters.** What is a recommended criterion to determine the number of clusters in a k-means approach? LO 4
   a. Silhouette score.
   b. Elbow method.
   c. Corresponding set of k clusters has an intuitive interpretation and does not change when input data is slightly changed.
   d. All of these.
7. **Properties of Single Linkage Agglomeration.** Which of the following statements about the single linkage method is false? LO 5
   a. In two dimensions, resulting clusters tend to be compact and circular.
   b. In two dimensions, the resulting clusters can be long and elongated.
   c. Chooses to merge the two clusters that have the most similar single pair of observations between the two clusters.
   d. May merge two clusters so that the resulting cluster has many observations dramatically different from each other.
8. **Comparing Agglomeration Methods.** If one wants to attempt to keep the largest difference between two observations in a cluster as small as possible, which linkage method may be most appropriate? LO 5
   a. Complete linkage
   b. Single linkage
   c. Ward's linkage
   d. Group average linkage
9. **Evaluating Association Rules.** When creating association rules, why is it common to only consider rules with a confidence measure above a minimum threshold? LO 7
   a. A minimum confidence threshold is not necessary if item granularity is appropriately defined.
   b. Rules with too low of a confidence measure do not occur frequently enough to be useful for business decision making.
   c. Rules with too low of a confidence measure are not surprising.
   d. Rules with too low of a confidence measure are unlikely to be true.
10. **Item Granularity in Transactional Data.** In a transactional data, what is the effect of aggregating similar items into a broader product category? LO 7
    a. Item granularity does not matter as one can simply adjust the minimum support threshold.
    b. The amount of insight derived from an association rule using the product category increases.
    c. Aggregation should be avoided and virtual items should be used instead.
    d. The support of the resulting product category increases.

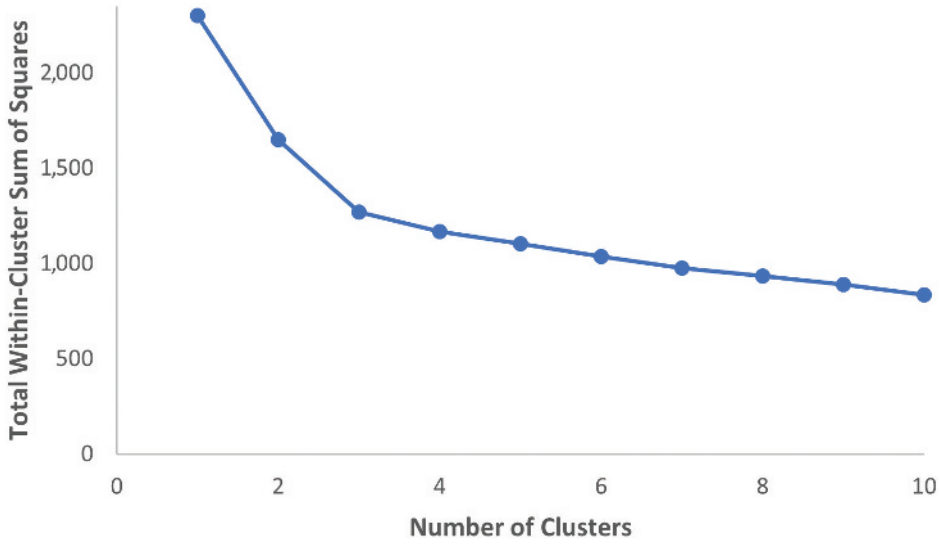
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

11. **Properties of Lift Ratio.** Which of the following statements about lift is false? LO 7
    a. Lift can be interpreted as how much more likely the antecedent and consequent itemsets occur together than if they were totally unrelated.
    b. Lift captures how surprising a rule is.
    c. Rules with high confidence can have low lift.
    d. Rules with low lift can provide insight on marketing opportunities.
12. **Support of Association Rules.** When creating association rules, why is it common to only consider rules with a support measure above a minimum threshold? LO 7
    a. Rules with too low of a support are unlikely to be true.
    b. Rules with too low of a support measure may not occur frequently enough to represent true correlated itemsets or to be useful for business decision making.
    c. Rules with too low of a support are unsurprising.
    d. A minimum support threshold is unnecessary if item granularity is defined appropriately.
13. **Virtual Items in Association Rules.** In transactional data, what are virtual items? LO 7
    a. Items involved in association rules with a high lift measure.
    b. Items corresponding to transactions occuring in a virtual reality setting.
    c. Product groupings resulting from the aggregation of similar items.
    d. Items not explicity part of the transaction, but characteristics associated with the transaction, e.g., location, time.
14. **Term Frequency Times Inverse Document Frequency (TFIDF).** What is a benefit of using TFIDF instead of just term frequency in text mining? LO 8
    a. TFIDF can be used in a bag-of-words approach, but term frequency cannot.
    b. TFIDF is a normalized measure that accounts for the varying lengths of documents in a corpus, but term frequency is not.
    c. TFIDF considers how often a term occurs in a document as well as how rare it is in the other documents of the corpus while term frequency only considers how often a term occurs.
    d. TFIDF can be used in word clouds, but term frequency cannot.
15. **Distance Measure for Text Mining.** Which of the following distance measures is commonly applied to a frequency-document matrix and why? LO 2, 3, 8
    a. Jaccard distance—because text is expressed as binary variables in a frequency-document matrix.
    b. Manhattan distance—because text is expressed quantitatively in a frequency-document matrix and it avoids the effect of outliers.
    c. Euclidean distance—because text is expressed quantitatively in a frequency-document matrix and outliers have been filtered out by upper and lower limits on term frequencies.
    d. Cosine distance—because it identifies similarity in term usage patterns instead of the magnitudes in term frequency measures.
16. **Preprocessing in Text Mining.** Which of the following is NOT usually included in the tokenization of text? LO 8
    a. Removal of punctuation and symbols
    b. Generation of a word cloud
    c. Removal of stopwords
    d. Stemming words
17. ***k*-Means Clustering of Wines.** The Boleyn sisters, a pair of entrepreneurs who recently sold their start-up for a multi-million-dollar sum, are looking for alternate investments for their newfound fortune. They are considering an investment in wine, similar to how some people invest in rare coins and fine art. To learn more about the properties of fine wine, they have collected data on 13 different characteristics of 178 wines. They have applied *k*-means clustering to these data for a range of *k* values and generated the following table of average silhouette values and elbow chart. LO 4

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Which value of k appears to be the most appropriate to categorize these wines?
b. Because the average silhouette values are low for all values of k, the Boleyn sisters are concerned about the quality of their clustering results. What experiments could they conduct to potentially improve the clusters?

k	Average Silhouette Value
2	0.259
3	0.285
4	0.260
5	0.231
6	0.194
7	0.206
8	0.197
9	0.156
10	0.202



DATA*file*
gatsby

18. **Distance to Centroid Calculation for Wine Clusters.** Jay Gatsby categorizes wines into one of three clusters. The centroids of these clusters (in standardized units), describing the average characteristics of a wine in each cluster, are listed in the following table.

Characteristic	Cluster 1	Cluster 2	Cluster 3
Alcohol	0.819	0.164	−0.937
Malic Acid	−0.329	0.869	−0.368
Ash	0.248	0.186	−0.393
Alkalinity	−0.677	0.523	0.249
Magnesium	0.643	−0.075	−0.573
Phenols	0.825	0.977	−0.034
Flavonoids	0.896	−1.212	0.083

(*continued*)

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Characteristic	Cluster 1	Cluster 2	Cluster 3
Nonflavonoids	−0.595	0.724	0.009
Proanthocyanidins	0.619	−0.778	0.010
Color Intensity	0.135	0.939	−0.881
Hue	0.497	−1.162	0.437
Dilution	0.744	−1.289	0.295
Proline	1.117	−0.406	−0.776

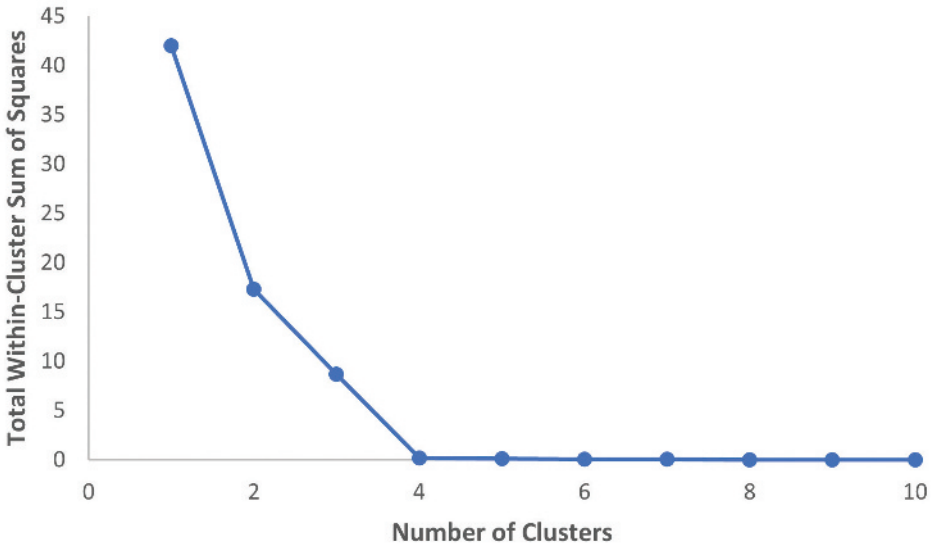
Jay has recently discovered a new wine from the Piedmont region of Italy with the following characteristics. In which cluster of wines should Jay place this new wine? Justify your choice with appropriate calculations. **LO 2**

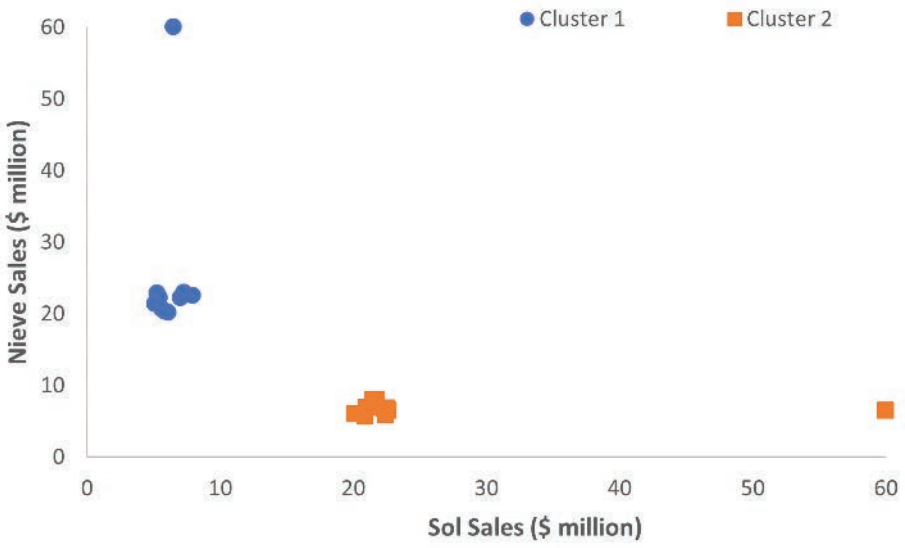
Characteristic	
Alcohol	−1.023
Malic Acid	−0.480
Ash	0.049
Alkalinity	0.600
Magnesium	−1.242
Phenols	1.094
Flavonoids	0.001
Nonflavonoids	0.548
Proanthocyanidins	−0.229
Color Intensity	−0.797
Hue	0.711
Dilution	−0.425
Proline	0.010

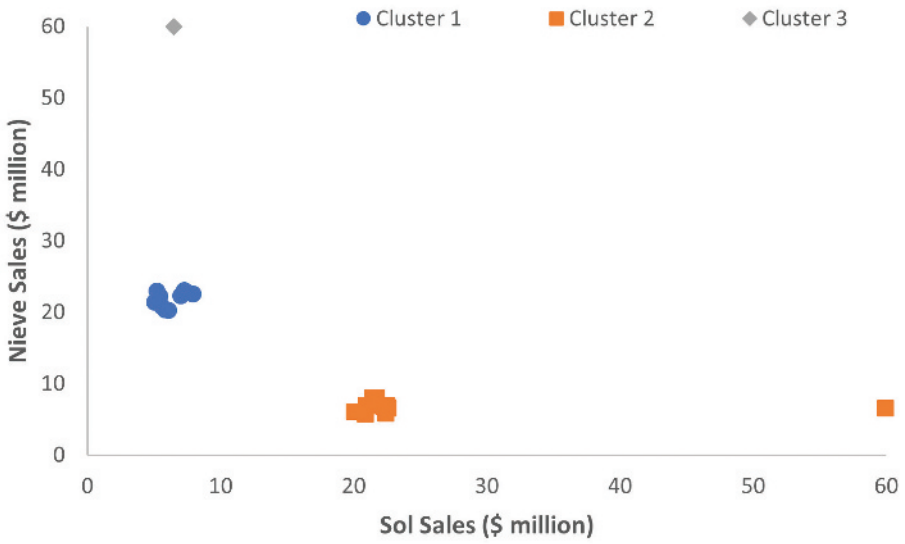
19. **Outliers' Impact on Clustering.** Sol & Nieve is a sporting good and outdoor gear retailer that operates in North America and Central America. In an attempt to characterize its stores and reassess Sol & Nieve's supply chain operations, Gustavo Esposito is analyzing sales in 22 regions for its two primary product lines: sol (beach-oriented apparel) and nieve (mountain-oriented apparel). For a range of k values, Gustavo has generated the following table of average silhouette values and elbow chart. After examining this output, Gustavo also generated the following scatter charts for the clusters correspond to $k = 2$, 3, and 4. **LO 4**
    a. What value of k is the most appropriate for these data?
    b. For the value of k identified in part (a), how should Gustavo interpret the results to characterize the sales regions?

k	Average Silhouette Value
2	0.741
3	0.779
4	0.837
5	0.667
6	0.401
7	0.433
8	0.502
9	0.536
10	0.494

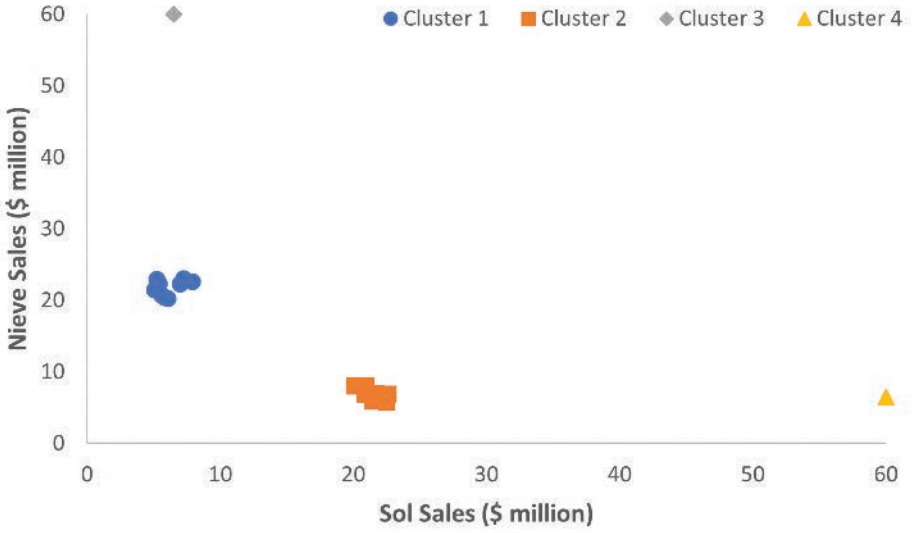
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.







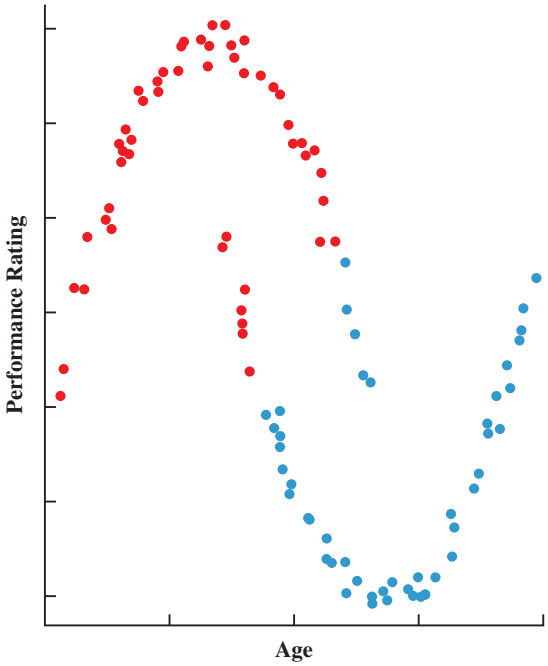
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



20. **Cluster Shapes for *k*-Means versus Single Linkage.** Heidi Zahn is a human resources manager currently reviewing data on 98 employees. In the data, each observation consists of an employee's age and an employee's performance rating. **LO 4, 5, 6**

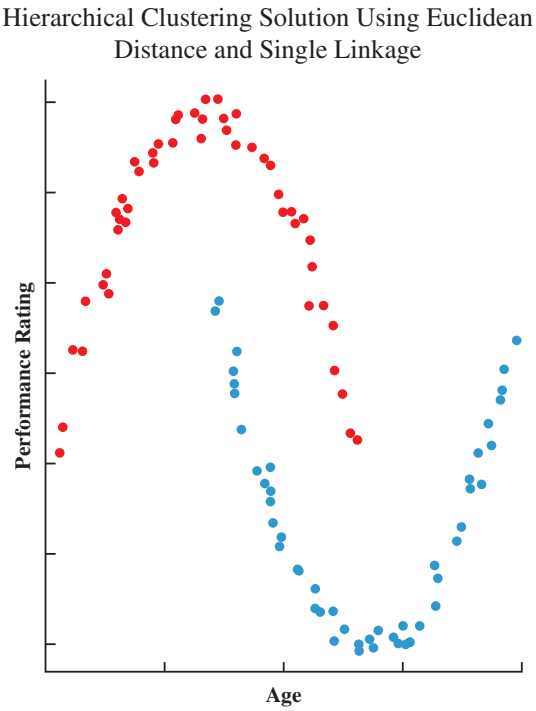
a. Heidi applied *k*-means clustering with $k = 2$ to the data and generated the following plot to visualize the clusters. Based on this plot, qualitatively characterize the two clusters of employees categorized by the *k*-means approach.

k-Means Clustering Solution with $k = 2$



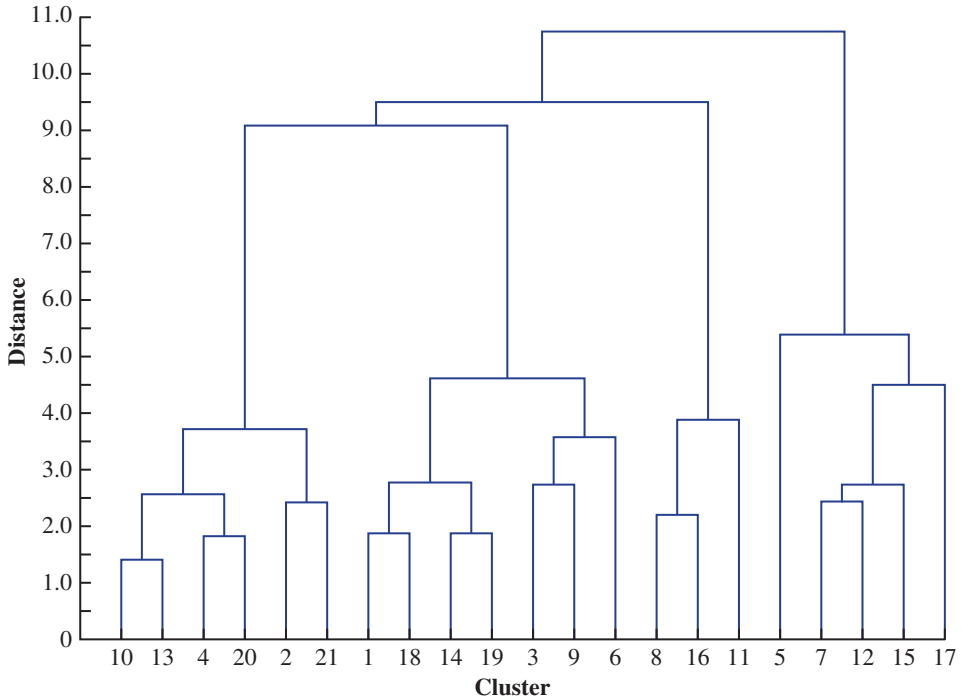
b. For a comparison, Heidi applied hierarchical clustering with the Euclidean distance and single linkage to the data and generated the following plot based on the level of agglomeration with two clusters. Based on this plot, qualitatively characterize the two clusters of employees categorized by the hierarchical approach.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



c. Which of the two approaches (*k*-means clustering from part (a) or hierarchical clustering from part (b)) would you recommend?

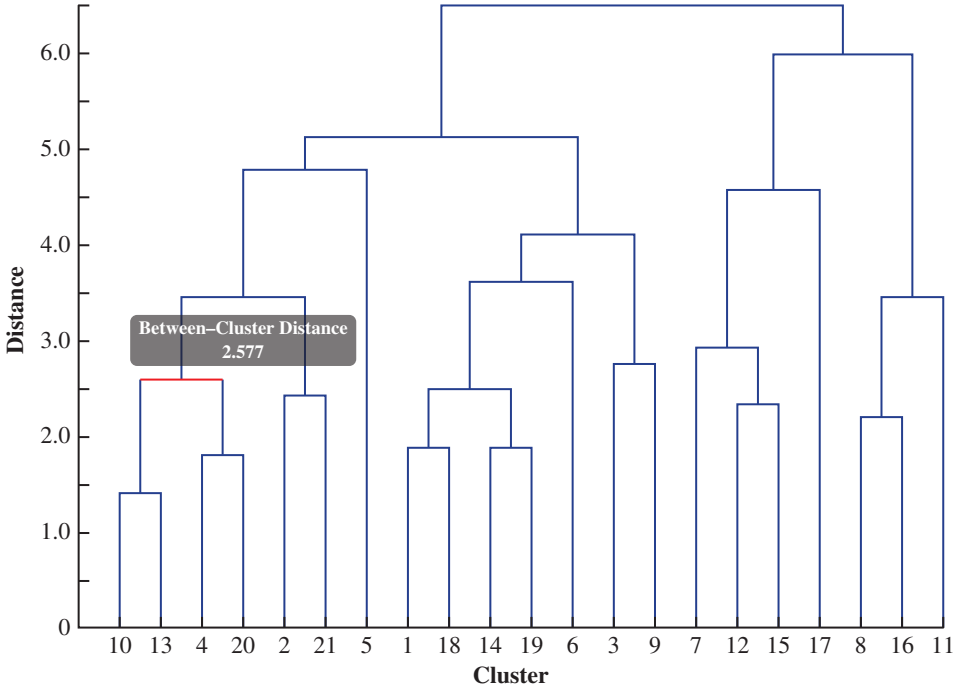
21. **Dendrogram of Utility Companies.** The regulation of electric and gas utilities is an important public policy question affecting consumer's choice and cost of energy provider. To inform deliberation on public policy, data on eight quantitative variables have been collected for a group of energy companies. To summarize the data, hierarchical clustering has been executed using Euclidean distance to measure dissimilarity between observations and Ward's linkage as the agglomeration method. Based on the following dendrogram, what is the most appropriate number of clusters to organize these utility companies? LO 6



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

22. **Complete Linkage Clustering of Utility Companies.** In an effort to inform political leaders and economists discussing the deregulation of electric and gas utilities, data on eight quantitative variables from utility companies have been grouped using hierarchical clustering based on Euclidean distance to measure dissimilarity between observations and complete linkage as the agglomeration method. **LO 5, 6**

a. Based on the following dendrogram, what is the most appropriate number of clusters to organize these utility companies?



b. Using the following data on the Observations 10, 13, 4, and 20, confirm that the complete linkage distance between the cluster containing {10, 13} and the cluster containing {4, 20} is 2.577 units as displayed in the dendrogram.

DATA *file* utility

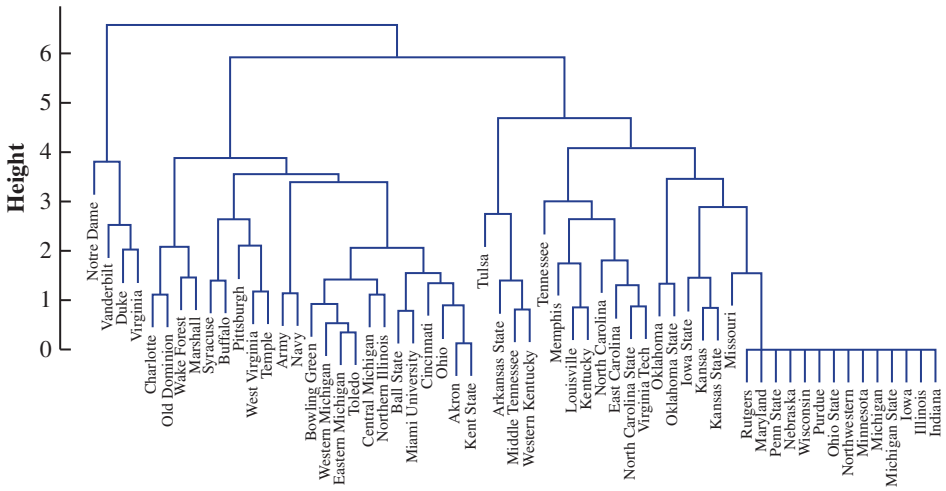
	Observation			
	10	13	4	20
Income/Debt	0.032	0.195	−0.510	0.466
Return	0.741	0.875	0.207	0.474
Cost	0.700	0.748	−0.004	−0.490
Load	−0.892	−0.735	−0.219	0.655
Peak	−0.173	1.013	−0.943	0.083
Sales	−0.693	−0.489	−0.702	−0.458
Percent Nuclear	1.620	2.275	1.328	1.733
Total Fuel Costs	−0.863	−1.035	−0.724	−0.721

23. **Interpreting Merge Steps from a Dendrogram.** From 1946 to 1990, the Big Ten Conference consisted of the University of Illinois, Indiana University, University of Iowa, University of Michigan, Michigan State University, University of Minnesota, Northwestern University, Ohio State University, Purdue University, and University of Wisconsin. In 1990, the conference added Pennsylvania State University. In 2011, the conference added the University of Nebraska. In 2014, the University of Maryland and Rutgers University were added to the conference with speculation of more schools being added in the future.

Based on the football stadium capacity, latitude, longitude, and enrollment, the Big Ten commissioner is curious how a clustering algorithm would suggest the conference expand beyond its current 14 members. To represent the 14 member

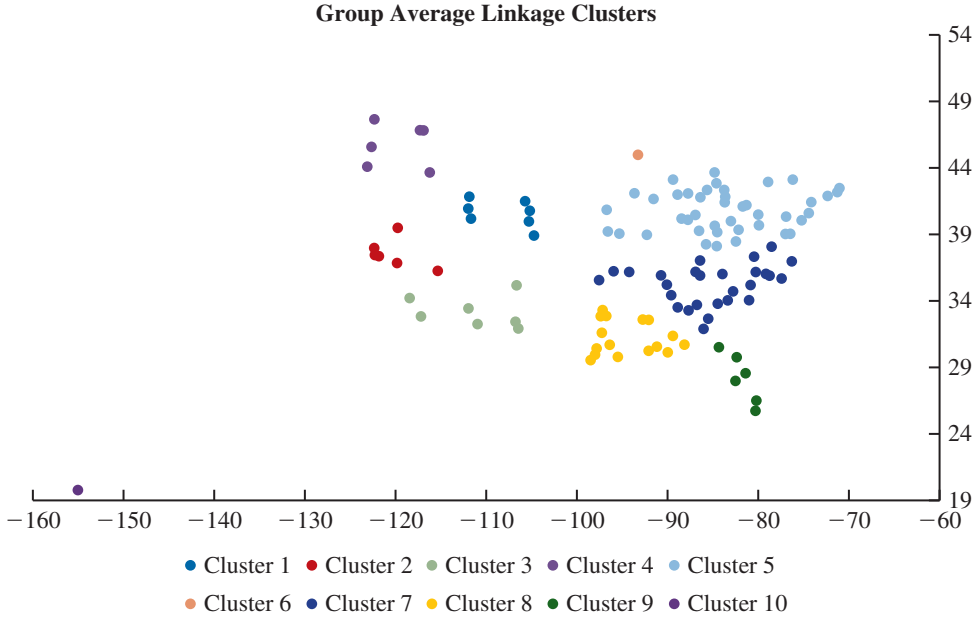
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

schools as an entity, each variable value for the 14 schools of the Big Ten Conference has been replaced with the respective variable median over these 14 schools. Using Euclidean distance to measure dissimilarity between observations, hierarchical clustering with complete linkage generates the following dendrogram. Describe the next three stages of conference expansion plan suggested by the dendrogram. **LO 6**



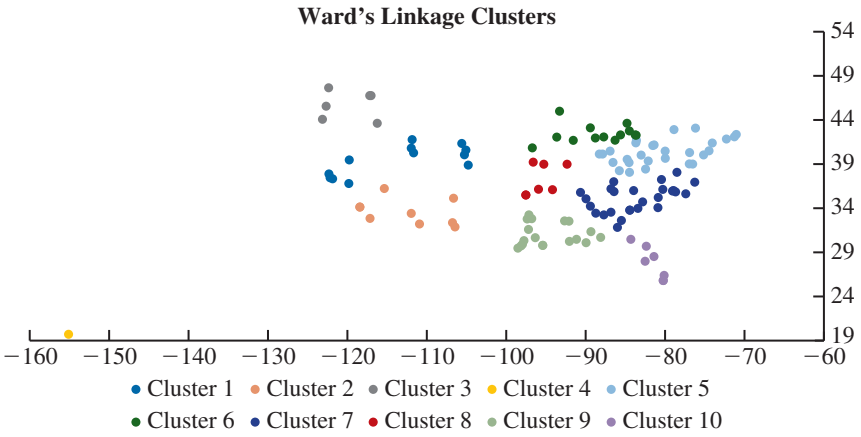
24. **Comparing Different Linkage Methods.** The Football Bowl Subdivision (FBS) level of the National Collegiate Athletic Association (NCAA) consists of over 100 schools. Most of these schools belong to one of several conferences, or collections of schools, that compete with each other on a regular basis in collegiate sports. Suppose the NCAA has commissioned a study that will propose the formation of conferences based on the similarities of constituent schools. If the NCAA cares only about geographic proximity of schools when determining conferences, it could use hierarchical clustering based on the schools' latitude and longitude values and Euclidean distance to compute dissimilarity between observations. The following charts and tables illustrate the 10-cluster solutions when using various linkage methods (group average, Ward's, complete, and single) to determine clusters of schools to be assigned to conferences using hierarchical clustering. Compare and contrast the resulting clusters created using each of these different linkage methods for hierarchical clustering. **LO 5, 6**

Group Average Linkage Clusters



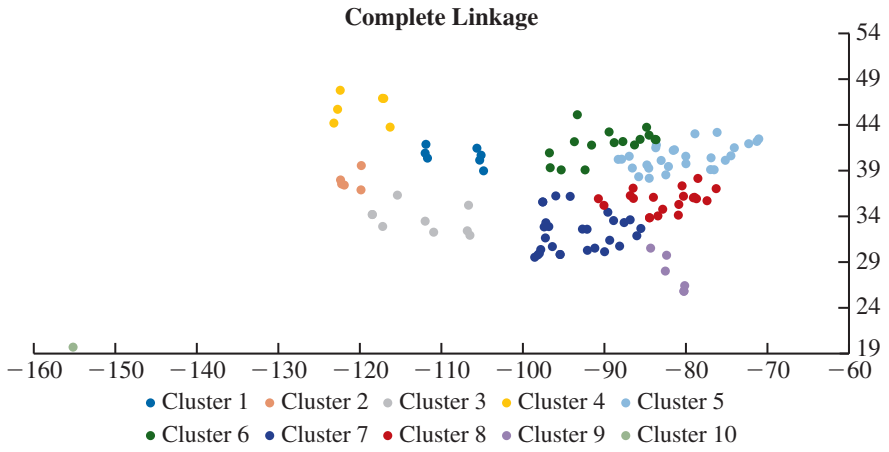
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Group Average Linkage Clusters	Count of School	Min of Latitude	Max of Latitude	Min of Longitude	Max of Longitude
1	7	38.9	41.7	−111.9	−104.8
2	6	36.2	39.4	−122.3	−115.3
3	8	31.8	35.1	−118.4	−106.4
4	6	43.6	47.6	−123.1	−116.2
5	43	38.0	43.6	−96.7	−71.0
6	1	45.0	45.0	−93.3	−93.3
7	30	31.8	38.0	−97.5	−76.2
8	18	29.5	33.2	−98.5	−88.1
9	7	25.8	30.5	−84.3	−80.1
10	1	19.7	19.7	−155.1	−155.1
Grand Total	**127**	**19.7**	**47.6**	**−155.1**	**−71.0**

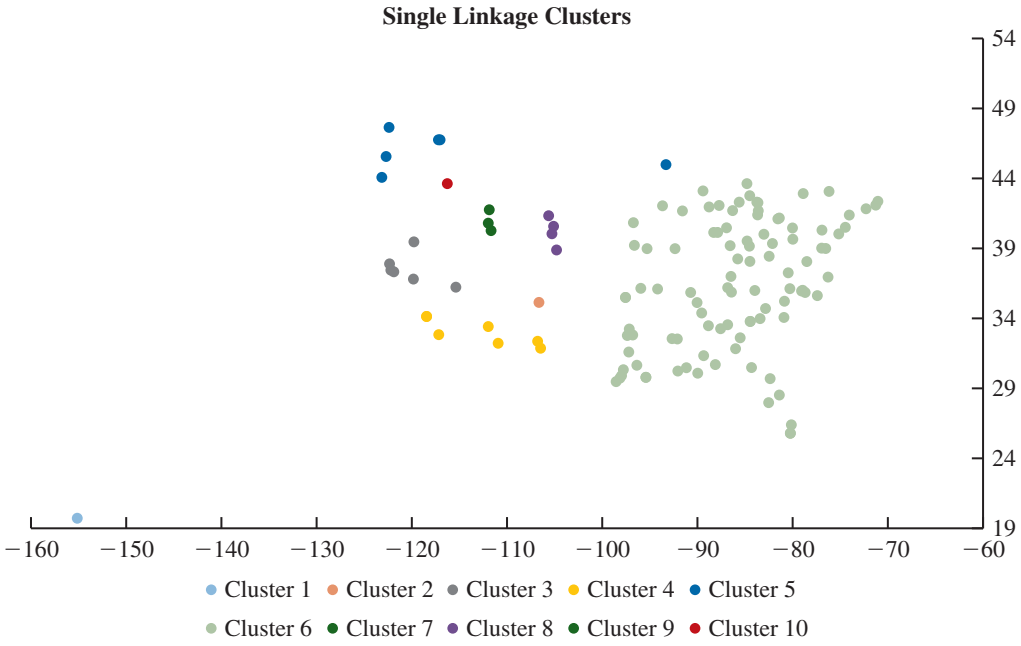


Ward's Linkage Clusters	Count of School	Min of Latitude	Max of Latitude	Min of Longitude	Max of Longitude
1	12	36.8	41.7	122.3	104.8
2	9	31.8	36.2	−118.4	−106.4
3	6	43.6	47.6	−123.1	−116.2
4	1	19.7	19.7	−155.1	−155.1
5	28	38.0	43.0	−88.3	−71.0
6	13	10.8	45.0	−96.7	−83.6
7	20	31.8	38.0	−90.7	−70.2
8	7	35.5	39.2	−97.5	−92.3
9	18	29.5	33.2	−98.5	−88.1
10	7	25.8	30.5	−84.3	−80.1
Grand Total	**127**	**19.7**	**47.6**	**−155.1**	**−71.0**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Complete Linkage Clusters	Count of School	Min of Latitude	Max of Latitude	Min of Longitude	Max of Longitude
1	7	38.9	41.7	−111.9	−104.8
2	5	36.8	39.2	−122.3	−119.7
3	9	31.8	36.2	−118.4	−106.4
4	6	43.6	47.6	−123.1	−116.2
5	28	38.0	43.0	−88.3	−71.0
6	16	39.0	45.0	−96.7	−83.6
7	28	29.5	36.1	−98.5	−85.5
8	20	33.8	38.0	−90.7	−76.2
9	7	25.8	30.5	−84.3	−80.1
10	1	19.7	19.7	−155.1	−155.1
Grand Total	**127**	**19.7**	**47.6**	**−155.1**	**−71.0**



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Single Linkage Clusters	Count of School	Min of Latitude	Max of Latitude	Min of Longitude	Max of Longitude
1	1	19.7	19.7	− 155.1	− 155.1
2	1	35.1	35.1	− 106.6	− 106.6
3	6	36.2	39.4	− 122.3	− 115.3
4	7	31.8	34.1	− 118.4	− 106.4
5	1	45.0	45.0	− 93.3	− 93.3
6	98	25.8	43.6	− 98.5	− 71.0
7	5	44.1	47.6	− 123.1	− 117.0
8	1	43.6	43.6	− 116.2	− 116.2
9	3	40.2	41.7	− 111.9	− 111.6
10	4	38.9	41.3	− 105.6	− 104.8
Grand Total	**127**	**19.7**	**47.6**	**− 155.1**	**− 71.0**

25. **Association Rules for Bookstore Transactions.** Leggere, an Internet book retailer, is interested in better understanding the purchase decisions of its customers. For a set of 1,604 customer transactions, it has categorized the individual book purchases comprising those transactions into one or more of the following categories: Novels, Willa Bean series, Cooking Books, Bob Villa Do-It-Yourself, Youth Fantasy, Art Books, Biography, Cooking Books by Mossimo Bottura, Harry Potter series, Florence Art Books, and Titian Art Books. Leggere has conducted association rules analysis on this data set and would like to analyze the output. Based on a minimum support of 10% and a minimum confidence of 50%, the table below shows the top 10 rules with respect to lift ratio. **LO 7**
    a. Explain why the fourth rule on the list is not helpful although it has 100% confidence and lift ratio of 1.861.
    b. Explain how the confidence of 55.6% and lift ratio of 1.851 was computed for the rule "If a customer buys a novel, a cooking book, and a biography book, then they buy an art book." Interpret these quantities.
    c. Based on these top 10 rules, what general insight can Leggere gain on the purchase habits of these customers?
    d. What will be the effect on the rules generated if Leggere decreases the minimum support and reruns the association rules analysis?
    e. What will be the effect on the rules generated if Leggere decreases the minimum confidence and reruns the association rules analysis?

Antecedent	Consequent	Support of A	Support of C	Support of A & C	Confidence	Lift Ratio
Novels, Cooking, Art	Biography	0.1552	0.3454	0.1016	0.655	1.895
Novels, Cooking, BobVilla	Art	0.1827	0.3005	0.1022	0.560	1.863
Novels, BotturaCooking	Cooking	0.1060	0.5374	0.1060	1.000	1.861
BotturaCooking	Cooking	0.1415	0.5374	0.1415	1.000	1.861
Novels, Cooking, Biography	Art	0.1827	0.3005	0.1016	0.556	1.851
Novels, Cooking, Art	BobVilla	0.1552	0.3635	0.1022	0.659	1.812
Cooking, BobVilla	Art	0.2363	0.3005	0.1278	0.541	1.800
Novels, Art	Biography	0.2007	0.3454	0.1241	0.618	1.789
Novels, YouthFantasy	Biography	0.1895	0.3454	0.1166	0.615	1.781
Cooking, Art	Biography	0.2082	0.3454	0.1272	0.611	1.768

26. **Congressional Voting Records.** Freelance reporter Irwin Fletcher is examining the historical voting records of members of the U.S. Congress. For 175 representatives, Irwin has collected the voting record (yes or no) on 16 pieces of legislation. To examine the relationship between representatives' votes on different issues, Irwin has

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

conducted an association rules analysis with a minimum support of 40% and a minimum confidence of 90%.

The data included the following bills:

Budget: approve federal budget resolution

Contras: aid for Nicaraguan contra rebels

El_Salvador: aid to El Salvador

Missile: funding for M-X missile program

Physician: freeze physician fees

Religious: equal access to all religious groups at schools

Satellite: ban on anti-satellite weapons testing

The following table shows the top five rules with respect to lift ratio. The table displays representatives' decisions in a "bill-vote" format. For example, "Contras-y" indicates that the representative voted yes on a bill to support the Nicaraguan Contra rebels and "Physician-n" indicates a no vote on a bill to freeze physician fees. **LO 7**

Antecedent	Consequent	Support of A&C	Confidence	Lift Ratio
Contras-y, Physician-n, Satellite-y	El_Salvador-n	0.40	0.95	1.98
Contras-y, Missile-y	El_Salvador-n	0.40	0.91	1.91
Contras-y, Physician-n	El_Salvador-n	0.44	0.91	1.90
Missile-n, Religious-y	El_Salvador-y	0.40	0.93	1.90
Budget-y, Contras-y, Physician-n	El_Salvador-n	0.41	0.90	1.89

a. Interpret the lift ratio of the first rule in the table.
b. What is the probability that a representative votes no on El Salvador aid given that they vote yes to aid to Nicaraguan Contra rebels and yes to the M-X missile program?
c. What is the probability that a representative votes no on El Salvador aid given that they vote no to the M-X missile program and yes to equal access to religious groups in schools?
d. What is the probability that a randomly selected representative votes yes on El Salvador aid?

Software Application

Problems 27 through 41 and Case Problem 2 require the use of data mining software (such as R or Orange). Guidelines for solving these problems with R and Orange are provided in the respective appendixes available in the Cengage eBook.

DATA*file*
breakfast.csv

27. **Breakfast Cereals.** Porridge Foods is a health-foods company considering a venture in the breakfast foods market. As part of its market research, Porridge Foods has collected data on 11 quantitative variables for 77 breakfast cereals contained in the file *breakfast*. Porridge Foods is interested in applying principal components analysis to summarize the information in these data. The following table lists the 11 quantitative variables. **LO 1**

Variable	Description
Calories	Calories per serving
Protein	Grams of protein
Fat	Grams of fat
Sodium	Milligrams of sodium
Fiber	Grams of dietary fiber
Carbo	Grams of complex carbohydrates
Sugars	Grams of sugars
Potass	Milligrams of potassium
Vitamins	Percentage of FDA recommended vitamins and minerals
Weight	Weight in ounces of one serving
Cups	Number of cups in one serving
Rating	Consumer reports rating of the cereal

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. What percentage of the variation is explained by the first principal component?
b. What percentage of variation is explained by the second principal component?
c. Provide interpretations for how the first principal component and the second principal component respectively describe the breakfast cereals. Support your narrative with observations from a combined analysis of the scatter plot of the principal components and the weights of the original variables in each of the principal components.

28. **Congressional Voting.** The file *congress* contains the voting record sample from the U.S. House of Representatives. These data include the voting record for 445 representatives on 1647 issues with votes recorded as −1 for nay, +1 for yay, and 0 for abstain (or absent).[1] Perform principal components analysis on the 1647 voting issues to summarize the information it may contain. **LO 1**

DATA *file*
congress.csv

a. What percentage of the variation is explained by the first principal component?
b. What percentage of the variation is explained by the second principal component?
c. Generate a scatter plot with the first principal component on the horizontal axis and the second principal component on the vertical axis.
d. Analyze and interpret the first principal component to identify how it distinguishes congress members. Based on the first principal component, what explains a large amount of the variation in how congress members vote?
e. Analyze and interpret the second principal component to identify how it distinguishes congress members.

29. ***k*-Means Clustering of Employees.** IBM employs a network of expert analytics consultants for various projects. To help it determine how to distribute its bonuses, IBM wants to form groups of employees with similar performance according to key performance metrics. Each observation (corresponding to an employee) in the file *ibm* consists of values for UsageRate which corresponds to the proportion of time that the employee has been actively working on high-priority projects, Recognition which is the number of projects for which the employee was specifically requested, and Leader which is the number of projects on which the employee has served as project leader. Apply *k*-means clustering with values of $k = 2$ to 8. Standardize the values of the input variables to adjust for the different magnitudes of the variables. How many clusters do you recommend for categorizing the employees? Why? **LO 2, 4**

DATA *file*
ibm.csv

30. ***k*-Means Clustering of Adam Sandler Movies.** Attracted by the possible returns from a portfolio of movies, hedge funds have invested in the movie industry by financially backing individual films and/or studios. The hedge fund Star Ventures is currently conducting some research involving movies involving Adam Sandler, an American actor, screenwriter, and film producer. As a first step, Star Ventures would like to cluster Adam Sandler movies based on their gross box office returns and movie critic ratings. Using the data in the file *adam*, apply *k*-means clustering with $k = 3$ to characterize three different types of Adam Sandler movies. Base the clusters on the variables Rating and Box. Rating corresponds to movie ratings provided by critics (a higher score represents a movie receiving better reviews). Box represents the gross box office earnings in 2015 dollars. Standardize the values of the input variables to adjust for the different magnitudes of the variables. Report the characteristics of each cluster including a count of movies, the average rating of movies, and the average box office earnings of movies in each cluster. How would you describe the movies in each cluster? **LO 2, 4**

DATA *file*
adam.csv

31. ***k*-Means Clustering of Trader Joe's Stores.** Josephine Mater works for the supply-chain analytics division of Trader Joe's, a national chain of specialty grocery stores. Trader Joe's is considering a redesign of its supply chain. Josephine knows that Trader Joe's uses frequent truck shipments from its distribution centers to its retail stores. To keep transportation costs low, distribution centers are typically located near retail stores. The file *traderjoe* contains data on the location of Trader Joe's retail stores.

DATA *file*
traderjoe.csv

[1]Note that these 1647 variables are categorical (nay, yay, abstain/absent) and not quantitative. However, because all these variables are categorical (with the same set of three categories), then we can choose an arbitrary (but consistent) numeric encoding of these variables and apply principal component analysis.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Josephine would like to use k-means clustering to help guide a proposal for positioning distribution centers. LO 2, 4

a. Apply k-means clustering with $k = 2$ to 10 and identify the value of k that results in the largest silhouette score.
b. Explain why the value of k from part (a) may not necessarily correspond to the best number of distribution centers to establish for the supply chain. Does clustering still provide value for this supply chain planning?

32. **Clustering Colleges with *k*-Means.** The Football Bowl Subdivision (FBS) level of the National Collegiate Athletic Association (NCAA) consists of over 100 schools. Most of these schools belong to one of several conferences, or collections of schools, that compete with each other on a regular basis in collegiate sports. Suppose the NCAA has commissioned a study that will propose the formation of conferences based on the similarities of the constituent schools. The file *fbs* contains data on schools that belong to the Football Bowl Subdivision. Each row in this file contains information on a school. The variables include football stadium capacity, latitude, longitude, endowment, and undergraduate enrollment. LO 2, 4

DATA*file* fbs.csv

a. Apply k-means clustering with $k = 10$ using football stadium capacity, latitude, longitude, endowment, and enrollment as variables. Standardize the input variables to adjust for the different magnitudes of the variables. Consider the cluster containing the most schools and characterize it with respect to its cluster centroid.
b. What problems do you see with the plan for defining the school membership of the 10 conferences directly with the 10 clusters in part (a)?
c. Repeat part (a), but this time do not standardize the values of the input variables. What value of k results in the largest silhouette score? To understand how the clustering differs from part (a), identify the dominating factor(s) in the formation of these new clusters.

33. **Grouping Colleges with Hierarchical Clustering.** Refer to the clustering problem involving the file *fbs* described in Problem 32. Using Euclidean distance to compute dissimilarity between observations, apply hierarchical clustering employing Ward's linkage with 10 clusters using football stadium capacity, latitude, longitude, endowment, and enrollment as variables. Standardize the values of the input variables to adjust for the different magnitudes of the variables. LO 2, 5, 6

DATA*file* fbs.csv

a. Identify the cluster with the largest average football stadium capacity. Using all the variables, how would you characterize this cluster?
b. Examine the smallest cluster. Why does this cluster have so few members?

34. **Cluster Comparison of Single Linkage to Group Average Linking.** Refer to the clustering problem involving the file *fbs* described in Problem 32. Using Euclidean distance to compute dissimilarity between observations, apply hierarchical clustering with 10 clusters using latitude and longitude as variables. LO 2, 5, 6

DATA*file* fbs.csv

a. Execute hierarchical clustering with single linkage and use a scatter chart to visualize the clusters with longitude as the x-variable and latitude as the y-variable.
b. Execute hierarchical clustering with group average linkage and use a scatter chart to visualize the clusters with longitude as the x-variable and latitude as the y-variable.
c. Compare the results of parts (a) and (b). Explain the differences.

35. **Hierarchical Clustering of National Parks.** As a guide in establishing standard management guidelines for similar national parks, the National Park Service is interested in clustering the data on 59 national parks in the file *parks*. Use the variables Area (park size in acres), Visitors (number of annual visitors), Latitude, and Longitude as the basis for applying hierarchical clustering to form four clusters. Apply hierarchical clustering with Euclidean distance and Ward's linkage. Standardize the values of the input variables to adjust for the different magnitudes of the variables. LO 3, 5, 6

DATA*file* parks.csv

a. Create a scatter chart to visualize the clusters with longitude as the x-variable and latitude as the y-variable.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Create a scatter chart to visualize the clusters with Area as the *x*-variable and Visitors as the *y*-variable.
c. Based on a visual inspection of the dendrogram, are four clusters a good choice? Provide a description of each of the clusters.

DATA*file*
whisky.csv

36. **Hierarchical Clustering of Whiskys.** Stuart Rankin, manager of the specialty store All Things Scottish, is interested in devising a way to recommend different types of Scotch whisky to customers based on their preferences among Scotches that they have tried and liked. The file *whisky* contains data on 69 binary variables over five general whisky attributes (color, nose, body, palate, and finish) for 109 different types of Scotch. Using Jaccard distance to compute dissimilarity between observations, generate the dendrogram resulting from hierarchical clustering with group average linkage to answer the following questions. **LO 5, 6**
a. What is the most similar pair of whiskys?
b. For a customer who likes Glenburgie, which three whiskys should Stuart recommend?
c. Which whisky appears to be the most unusual?

DATA*file*
apple_binary.csv
apple_stacked.csv

37. **iStore Transactions.** Apple Inc. tracks online transactions at its iStore and is interested in learning about the purchase patterns of its customers in order to provide recommendations as a customer browses its web site. A sample of the "shopping cart" data resides in the files *apple_binary* (binary matrix format) and *apple_stacked* (stacked data format). Use a minimum support of 10% of the total number of transactions and a minimum confidence of 50% to generate a list of association rules. **LO 7**

Variable	Description
Retina Display	iPad equipped with liquid crystal display
Cellular	iPad equipped with cellular connectivity
16GB	16 gigabytes of memory
32GB	32 gigabytes of memory
64GB	64 gigabytes of memory
Case	hard-shelled clam iPad case
Stand	stand to prop up iPad
Speakers	external plug-in speakers
Plug Bug	dual charger for iPad
Camera Kit	camera connection kit
Wireless Keyboard	bluetooth external keyboard
Stylus	fabric tip writing stylus

a. Interpret what the rule with the largest lift ratio is saying about the relationship between the antecedent itemset and consequent itemset.
b. Interpret the confidence of the rule with the largest lift ratio.
c. Interpret the lift ratio of the rule with the largest lift ratio.
d. Review the top 15 rules and summarize what the rules suggest.

DATA*file*
cookie_binary.csv
cookie_stacked.csv

38. **Browser Histories.** Cookie Monster Inc. is a company that specializes in the development of software that tracks web browsing history of individuals. Cookie Monster Inc. is interested in analyzing its data to gain insight on the online behavior of individuals. A sample of browser histories is provided in the files *cookie_binary* (binary matrix format) and *cookie_stacked* (stacked data format) that indicate which websites were visited by which customers. Use a minimum support of 4% of the transactions (800 of the 20,000 total transactions) and a minimum confidence of 50% to generate a list of association rules. **LO 7**
a. Which three web sites appear in the association rules with the largest lift ratio?
b. Identify the association rule with the largest lift ratio that also has Pinterest as the antecedent. What is the consequent web site in this rule?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Interpret the confidence of the rule from part (b). Provide a logical narrative speculating why the event described by this rule would occur.
d. Identify the association rule with the largest lift ratio that also has TheEveryGirl as the antecedent. What is the consequent web site in this rule?
e. Interpret the lift ratio of the rule from part (d).

39. **Grocery Store Transactions.** A grocery store introducing items from Italy is interested in analyzing buying trends of these new "international" items, namely prosciutto, Peroni, risotto, and gelato. The files *grocery_binary* (binary matrix format) and *grocery_stacked* (stacked data format) provide data on a collection of transactions. **LO 7**

DATA *file*
grocery_binary.csv
grocery_stacked.csv

a. Use a minimum support of 10% of the transactions (100 of the 1,000 total transactions) and a minimum confidence of 50% to generate a list of association rules. How many rules satisfy this criterion?
b. Use a minimum support of 25% of the transactions (250 of the 1,000 total transactions) and a minimum confidence of 50% to generate a list of association rules. How many rules satisfy this criterion? Why may the grocery store want to increase the minimum support required for their analysis? What is the risk of increasing the minimum support required?
c. Using the list of rules from part (b), consider the rule with the largest lift ratio that also involves an Italian item. Interpret what this rule is saying about the relationship between the antecedent itemset and consequent itemset.
d. Interpret the confidence of the rule with the largest lift ratio that also involves an Italian item.
e. Interpret the lift ratio of the rule with the largest lift ratio that also involves an Italian item.
f. What insight can the grocery store obtain about its purchasers of the Italian fare?

40. **Tweets.** Companies can learn a lot about customer experiences by monitoring the social media web site Twitter. The file *airlinetweets* contains a sample of 36 tweets of an airline's customers. Preprocess the text and generate frequency and binary document-term matrices. **LO 8**

DATA *file*
airlinetweets.csv

a. Which five terms appear in the most tweets? In how many tweets does each term appear?
b. Using Jaccard's distance to compute dissimilarity between observations, apply hierarchical clustering employing Ward's linkage to yield four clusters on the binary document-term matrix. How many documents are in each cluster? Give a description of each cluster.
c. How could management use the results obtained in part (b)?

41. **Yelp Reviews.** The online review service Yelp helps millions of consumers find the goods and services they seek. To help consumers make more-informed choices, Yelp includes over 120 million reviews. The file *yelp* contains a sample of 21 reviews for an Italian restaurant. Preprocess the text and generate frequency and binary document-term matrices. **LO 8**

DATA *file*
yelp.csv

a. Which five terms appear in the most reviews? In how many reviews does each term appear?
b. Using Jaccard's distance to compute dissimilarity between observations, apply hierarchical clustering employing Ward's linkage to yield two clusters from the binary document-term matrix from the reviews. How many documents are in each cluster? Give a description of each cluster.

Case Problem 1: Big Ten Expansion

Case Problem 1 does not require the use of data mining software, but the use of Excel is recommended.

From 1946 to 1990, the Big Ten Conference consisted of the University of Illinois, Indiana University, University of Iowa, University of Michigan, Michigan State University, University of Minnesota, Northwestern University, Ohio State University, Purdue University, and University of Wisconsin. In 1990, the conference added Pennsylvania State University.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In 2011, the conference added the University of Nebraska. In 2014, the University of Maryland and Rutgers University were added to the conference with speculation of more schools being added in the future.

DATA*file* big10expand

The file *big10expand* contains data on the football stadium capacity, latitude, longitude, endowment, and enrollment of 59 National Collegiate Athletic Association (NCAA) Football Bowl Subdivision (FBS) schools. Treat the 10 schools that were members of the Big Ten from 1946 to 1990 as being in a cluster and the other 49 schools as each being in their own cluster. **LO 2, 5, 6**

1. Using Euclidean distance to measure dissimilarity between observations, determine which school (in its own cluster of one) that hierarchical clustering with complete linkage would recommend integrating into the Big Ten Conference. That is, which school is the most similar with respect to complete linkage to the cluster of ten schools that were members of the Big Ten from 1946 to 1990?
2. Add the single school identified in part (1) to create a cluster of 11 schools representing a hypothetical Big Ten Conference. Repeat the calculations to identify the school most similar with respect to complete linkage to this new cluster of 11 schools.
3. Add the school identified in part (2) to create a cluster of 12 schools representing a hypothetical Big Ten Conference. Repeat the calculations to identify the school most similar with respect to complete linkage to this new cluster of 12 schools.
4. Add the school identified in part (3) to create a cluster of 13 schools representing a hypothetical Big Ten Conference. Repeat the calculations to identify the school most similar with respect to complete linkage to this new cluster of 13 schools. Add this school to create a 14-school cluster.
5. How does the hypothetical 14-team cluster created in part (4) compare to the actual 14-team Big Ten Conference? For both the hypothetical 14-team Big Ten Conference and the actual 14-team Big Ten Conference, compute the cluster centroid, the distance from each cluster member to the cluster centroid, and average distance between the observations in the cluster. What do you observe when comparing these two clusters? Which cluster has the smaller average distance between observations? Is this surprising? Explain.

Case Problem 2: Know Thy Customer

Case Problem 2 requires the use of data mining software such as R or Orange.

Know Thy Customer (KTC) is a financial consulting company that provides personalized financial advice to its clients. As a basis for developing this tailored advising, KTC would like to segment its customers into several representative groups based on key characteristics. Peyton Blake, the director of KTC's fledging analytics division, plans to establish the set of representative customer profiles based on 600 customer records in the file *ktc*. Each customer record contains data on age, annual income, marital status, number of children, level of education, and whether the customer rents or owns their home. KTC's market research staff has determined that these six characteristics should form the basis of the customer clustering.

Peyton has invited a summer intern, Danny Riles, into the office so they can discuss how to proceed. As they review the data on the computer screen, Peyton's brow furrows as they realize that this task may not be trivial. The data contains both categorical variables (Marital, Education, and Home) and quantitative variables (Age, Income, and Children). **LO 2, 5, 6**

DATA*file* ktc.csv

Managerial Report

Apply cluster analysis to establish a set of customer profiles. Let the following steps guide your analysis and structure your report.

1. To address these mixed data (observations containing both quantitative and categorical variables), derive a new set of a binary variables based on the original variables according to the following rules.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$$\text{Retirement} = \begin{cases} 1 \text{ if Age} > 62 \\ 0 \text{ otherwise} \end{cases}$$

$$\text{MiddleUpper} = \begin{cases} 1 \text{ if Income} > 85{,}000 \\ 0 \text{ otherwise} \end{cases}$$

$$\text{Married} = \begin{cases} 1 \text{ if Marital} = \text{“married”} \\ 0 \text{ otherwise} \end{cases}$$

$$\text{FamilyPlan} = \begin{cases} 1 \text{ if Children} > 0 \\ 0 \text{ otherwise} \end{cases}$$

$$\text{MastersPlus} = \begin{cases} 1 \text{ if Education} = \text{“masters” or “more than masters”} \\ 0 \text{ otherwise} \end{cases}$$

$$\text{Own} = \begin{cases} 1 \text{ if Home} = \text{"own"} \\ 0 \text{ otherwise} \end{cases}$$

2. Using the binary variables Retirement, MiddleUpper, Married, FamilyPlan, MastersPlus, and Own created in part (1), apply hierarchical clustering with an appropriate distance metric and linkage method. Identify the appropriate number of clusters and characterize these clusters.
3. Peyton would like to compare the clusters obtained in part (2) with a clustering in which all variables are treated as quantitative. Peyton recognizes that the Education variable is an ordinal variable because there is a natural ordering of the various levels of education. Therefore, Peyton decides to encode the Education variable numerically as:

$$\text{EdLevel} = \begin{cases} 1 \text{ if Education} = \text{“less than high school”} \\ 2 \text{ if Education} = \text{“high school”} \\ 3 \text{ if Education} = \text{“associates”} \\ 4 \text{ if Education} = \text{“bachelors”} \\ 5 \text{ if Education} = \text{“masters”} \\ 6 \text{ if Education} = \text{“more than masters”} \end{cases}$$

Using the variables Age, Income, Married, Children, EdLevel, and Own, apply *k*-means clustering various values of *k* to identify a recommended number of clusters. Characterize these clusters.

4. Compare the clusters from part (2) to those in part (3). What set of customer profiles do you recommend to KTC?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 7

Statistical Inference

Contents

Analytics in Action:
John Morrell & Company

7.1 Selecting a Sample
Sampling from a Finite Population
Sampling from an Infinite Population

7.2 Point Estimation
Practical Advice

7.3 Sampling Distributions
Sampling Distribution of $\bar{x}$
Sampling Distribution of $\bar{p}$

7.4 Interval Estimation
Interval Estimation of the Population Mean
Interval Estimation of the Population Proportion

7.5 Hypothesis Tests
Developing Null and Alternative Hypotheses
Type I and Type II Errors
Hypothesis Test of the Population Mean
Hypothesis Test of the Population Proportion

7.6 Big Data, Statistical Inference, and Practical Significance
Sampling Error
Nonsampling Error
Big Data
Understanding What Big Data Is
Big Data and Sampling Error
Big Data and the Precision of Confidence Intervals
Implications of Big Data for Confidence Intervals
Big Data, Hypothesis Testing, and *p* Values
Implications of Big Data in Hypothesis Testing

Summary 376
Glossary 377
Problems 380

Available in the Cengage eBook:
Appendix: Random Sampling with R
Appendix: Interval Estimation with R
Appendix: Hypothesis Testing with R

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Take random samples from finite and infinite populations.

LO 2 Calculate and interpret point estimates, margins of error, and confidence intervals for population means.

LO 3 Calculate and interpret point estimates, margins of error, and confidence intervals for population proportions.

LO 4 Calculate point estimates of population standard deviations.

LO 5 Identify the sampling distributions for sample means and sample proportions.

LO 6 Formulate and interpret appropriate one-tailed and two-tailed null and alternative hypotheses for problem scenarios.

LO 7 Test one-tailed and two-tailed hypotheses about a population mean and interpret the results.

LO 8 Test one-tailed and two-tailed hypotheses about a population proportion and interpret the results.

LO 9 Identify the Type I error and Type II error associated with a hypothesis test and explain their potential consequences.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

John Morrell & Company*

Cincinnati Ohio

John Morrell & Company, which was established in England in 1827, is considered the oldest continuously operating meat manufacturer in the United States. It is a wholly owned and independently managed subsidiary of Smithfield Foods, Smithfield, Virginia. John Morrell & Company offers an extensive product line of processed meats and fresh pork to consumers under 13 regional brands, including John Morrell, E-Z-Cut, Tobin's First Prize, Dinner Bell, Hunter, Kretschmar, Rath, Rodeo, Shenson, Farmers Hickory Brand, Iowa Quality, and Peyton's. Each regional brand enjoys high brand recognition and loyalty among consumers.

Market research at Morrell provides management with up-to-date information on the company's various products and how the products compare with competing brands of similar products. In order to compare a beef pot roast made by Morrell to similar beef products from two major competitors, Morrell asked a random sample of consumers to indicate how the products rated in terms of taste, appearance, aroma, and overall preference.

In Morrell's independent taste-test study, a sample of 224 consumers in Cincinnati, Milwaukee, and Los Angeles was chosen. Of these 224 consumers, 150 preferred the beef pot roast made by Morrell. Based on these results, Morrell estimates that the population proportion that prefers Morrell's beef pot roast is $\bar{p} = 150/224 = 0.67$. Recognizing that this estimate is subject to sampling error, Morrell calculates the 95% confidence interval for the population proportion that prefers Morrell's beef pot roast to be 0.6080 to 0.7312.

Morrell then turned its attention to whether these sample data support the conclusion that Morrell's beef pot roast is the preferred choice of more than 50% of the consumer population. Letting p indicate the proportion of the population that prefers Morrell's product, the hypothesis test for the research question is as follows:

$$H_0: p \leq 0.50$$
$$H_a: p > 0.50$$

The null hypothesis H_0 indicates the preference for Morrell's product is less than or equal to 50%. If the sample data support rejecting H_0 in favor of the alternative hypothesis H_a, Morrell will draw the research conclusion that in a three-product comparison, its beef pot roast is preferred by more than 50% of the consumer population. Using statistical hypothesis testing procedures, the null hypothesis H_0 was rejected. The study provided statistical evidence supporting H_a and the conclusion that the Morrell product is preferred by more than 50% of the consumer population.

In this chapter, you will learn about simple random sampling and the sample selection process. In addition, you will learn how statistics such as the sample mean and sample proportion are used to estimate parameters such as the population mean and population proportion. The concept of a sampling distribution will be introduced and used to compute the margins of error associated with sample estimates. You will then learn how to use this information to construct and interpret interval estimates of a population mean and a population proportion. We then discuss how to formulate hypotheses and how to conduct tests such as the one used by Morrell. You will learn how to use sample data to determine whether or not a hypothesis should be rejected.

*The authors are indebted to Marty Butler, Vice President of Marketing, John Morrell, for providing this Analytics in Action.

Refer to Chapter 2 for a fundamental overview of data and descriptive statistics.

When collecting data, we usually want to learn about some characteristic(s) of the population, the collection of all the elements of interest, from which we are collecting that data. In order to know about some characteristic of a population with certainty, we must collect data from every element in the population of interest; such an effort is referred to as a **census**. However, there are many potential difficulties associated with taking a census:

- A census may be expensive; if resources are limited, it may not be feasible to take a census.
- A census may be time consuming; if the data need be collected quickly, a census may not be suitable.
- A census may be misleading; if the population is changing quickly, by the time a census is completed the data may be obsolete.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- A census may be unnecessary; if perfect information about the characteristic(s) of the population of interest is not required, a census may be excessive.
- A census may be impractical; if observations are destructive, taking a census would destroy the population of interest.

In order to overcome the potential difficulties associated with taking a census, we may decide to take a sample (a subset of the population) and subsequently use the sample data we collect to make inferences and answer research questions about the population of interest. Therefore, the objective of sampling is to gather data from a subset of the population that is as similar as possible to the entire population, so that what we learn from the sample data accurately reflects what we want to understand about the entire population. When we use the sample data we have collected to make estimates of or draw conclusions about one or more characteristics of a population (the value of one or more parameters), we are using the process of **statistical inference**.

A sample that is similar to the population from which it has been drawn is said to be representative of the population.

Sampling is done in a wide variety of research settings. Let us begin our discussion of statistical inference by citing two examples in which sampling was used to answer a research question about a population.

1. Members of a political party in Texas are considering giving their support to a particular candidate for election to the U.S. Senate, and party leaders want to estimate the proportion of registered voters in the state that favor the candidate. A sample of 400 registered voters in Texas is selected, and 160 of those voters indicate a preference for the candidate. Thus, an estimate of proportion of the population of registered voters who favor the candidate is $160/400 = 0.40$.
2. A tire manufacturer is considering production of a new tire designed to provide an increase in lifetime mileage over the firm's current line of tires. To estimate the mean useful life of the new tires, the manufacturer produced a sample of 120 tires for testing. The test results provided a sample mean of 36,500 miles. Hence, an estimate of the mean useful life for the population of new tires is 36,500 miles.

A sample mean provides an estimate of a population mean, and a sample proportion provides an estimate of a population proportion. With estimates such as these, some estimation error can be expected. This chapter provides the basis for determining how large that error might be.

It is important to realize that sample results provide only *estimates* of the values of the corresponding population characteristics. We do not expect exactly 0.40, or 40%, of the population of registered voters to favor the candidate, nor do we expect the sample mean of 36,500 miles to exactly equal the mean lifetime mileage for the population of all new tires produced. The reason is simply that the sample contains only a portion of the population and cannot be expected to perfectly replicate the population. Some error, or deviation of the sample from the population, is to be expected. With proper sampling methods, the sample results will provide "good" estimates of the population parameters. But how good can we expect the sample results to be? Fortunately, statistical procedures are available for answering this question.

Let us define some of the terms used in sampling. The **sampled population** is the population from which the sample is drawn, and a **frame** is a list of the elements from which the sample will be selected. In the first example, the sampled population is all registered voters in Texas, and the frame is a list of all the registered voters. Because the number of registered voters in Texas is a finite number, the first example is an illustration of sampling from a finite population.

The sampled population for the tire mileage example is more difficult to define because the sample of 120 tires was obtained from a production process at a particular point in time. We can think of the sampled population as the conceptual population of all the tires that could have been made by the production process at that particular point in time. In this sense, the sampled population is considered infinite, making it impossible to construct a frame from which to draw the sample.

In this chapter, we show how simple random sampling can be used to select a sample from a finite population and we describe how a random sample can be taken from an infinite population that is generated by an ongoing process. We then discuss how data obtained from a sample can be used to compute estimates of a population mean, a population standard deviation, and a population proportion. In addition, we introduce the important concept of a sampling distribution. As we will show, knowledge of the appropriate sampling distribution enables us

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to make statements about how close the sample estimates are to the corresponding population parameters, to compute the margins of error associated with these sample estimates, and to construct and interpret interval estimates. We then discuss how to formulate hypotheses and how to use sample data to conduct tests of a population means and a population proportion.

7.1 Selecting a Sample

The director of personnel for Electronics Associates, Inc. (EAI) has been assigned the task of developing a profile of the company's 2,500 employees. The characteristics to be identified include the mean annual salary for the employees and the proportion of employees having completed the company's management training program.

DATA*file*
eai

Using the 2,500 employees as the population for this study, we can find the annual salary and the training program status for each individual by referring to the firm's personnel records. The data set containing this information for all 2,500 employees in the population is in the file *eai*.

A measurable factor that defines a characteristic of a population, process, or system is called a **parameter**. For EAI, the population mean annual salary μ, the population standard deviation of annual salaries σ, and the population proportion p of employees who completed the training program are of interest to us. Using the EAI data, we compute the population mean and the population standard deviation for the annual salary data.

In Chapter 2, we discuss the computation of the mean and standard deviation of a population.

$$\text{Population mean:} \quad \mu = \$71{,}800$$
$$\text{Population standard deviation:} \quad \sigma = \$4{,}000$$

The data for the training program status show that 1,500 of the 2,500 employees completed the training program. Letting p denote the proportion of the population that completed the training program, we see that $p = 1{,}500/2{,}500 = 0.60$. The population mean annual salary ($\mu = \$71{,}800$), the population standard deviation of annual salary ($\sigma = \$4{,}000$), and the population proportion that completed the training program ($p = 0.60$) are parameters of the population of EAI employees.

Often the cost of collecting information from a sample is substantially less than the cost of taking a census. Especially when personal interviews must be conducted to collect the information.

Now suppose that the necessary information on all the EAI employees was not readily available in the company's database. The question we must consider is how the firm's director of personnel can obtain estimates of the population parameters by using a sample of employees rather than all 2,500 employees in the population. Suppose that a sample of 30 employees will be used. Clearly, the time and the cost of developing a profile would be substantially less for 30 employees than for the entire population. If the personnel director could be assured that a sample of 30 employees would provide adequate information about the population of 2,500 employees, working with a sample would be preferable to working with the entire population. Let us explore the possibility of using a sample for the EAI study by first considering how we can identify a sample of 30 employees.

Sampling from a Finite Population

Statisticians recommend selecting a probability sample when sampling from a finite population because a probability sample allows you to make valid statistical inferences about the population. The simplest type of probability sample is one in which each sample of size n has the same probability of being selected. It is called a simple random sample. A simple random sample of size n from a finite population of size N is defined as follows.

Simple Random Sample (Finite Population)

A **simple random sample** of size n from a finite population of size N is a sample selected such that each possible sample of size n has the same probability of being selected.

The random numbers generated using Excel's RAND *function follow a uniform probability distribution between 0 and 1.*

Procedures used to select a simple random sample from a finite population are based on the use of random numbers. We can use Excel's RAND function to generate a random number between 0 and 1 by entering the formula *=RAND()* into any cell in a worksheet. The number generated is called a random number because the mathematical procedure used by the RAND

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

function guarantees that every number between 0 and 1 has the same probability of being selected. Let us see how these random numbers can be used to select a simple random sample.

Our procedure for selecting a simple random sample of size n from a population of size N involves two steps.

Excel's Sort *procedure is especially useful for identifying the* n *elements assigned the* n *smallest random numbers.*

Step 1. Assign a random number to each element of the population.
Step 2. Select the n elements corresponding to the n smallest random numbers.

Because each set of n elements in the population has the same probability of being assigned the n smallest random numbers, each set of n elements has the same probability of being selected for the sample. If we select the sample using this two-step procedure, every sample of size n has the same probability of being selected; thus, the sample selected satisfies the definition of a simple random sample.

Let us consider the process of selecting a simple random sample of 30 EAI employees from the population of 2,500. We begin by generating 2,500 random numbers, one for each employee in the population. Then we select 30 employees corresponding to the 30 smallest random numbers as our sample. Refer to Figure 7.1 as we describe the steps involved.

The random numbers generated by executing these steps will vary; therefore, results will not match Figure 7.1.

Step 1. In cell D1, enter the text *Random Numbers*
Step 2. In cells D2:D2501, enter the formula = *RAND()*
Step 3. Select the cell range D2:D2501
Step 4. In the **Home** tab in the Ribbon:
Click **Copy** in the **Clipboard** group
Click the arrow below **Paste** in the **Clipboard** group. When the **Paste** window appears, click **Values** in the **Paste Values** area
Press the **Esc** key
Step 5. Select cells A1:D2501
Step 6. In the **Data** tab on the Ribbon, click **Sort** in the **Sort & Filter** group
Step 7. When the **Sort** dialog box appears:
Select the check box for **My data has headers**
In the first **Sort by** dropdown menu, select **Random Numbers**
Click **OK**

After completing these steps we obtain a worksheet like the one shown on the right in Figure 7.1. The employees listed in rows 2–31 are the ones corresponding to the smallest 30 random numbers that were generated. Hence, this group of 30 employees is a simple random sample. Note that the random numbers shown on the right in Figure 7.1 are in ascending order, and that the employees are not in their original order. For instance, employee 1721 in the population is associated with the smallest random number and is the first element in the sample, and employee 27 in the population (see row 28 of the worksheet on the left) has been included as the 9th observation in the sample (row 10 of the worksheet on the right).

Sampling from an Infinite Population

Sometimes we want to select a sample from a population, but the population is infinitely large or the elements of the population are being generated by an ongoing process for which there is no limit on the number of elements that can be generated. Thus, it is not possible to develop a list of all the elements in the population. This is considered the infinite population case. With an infinite population, we cannot select a simple random sample because we cannot construct a frame consisting of all the elements. In the infinite population case, statisticians recommend selecting what is called a random sample.

Random Sample (Infinite Population)

A **random sample** of size n from an infinite population is a sample selected such that the following conditions are satisfied.

1. Each element selected comes from the same population.
2. Each element is selected independently.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.1 Using Excel to Select a Simple Random Sample

	A	B	C	D	E	F	G
1	Employee	Annual Salary	Training Program	Random Numbers			
2	1	75769.50	No	0.109886			
3	2	70823.00	Yes	0.217989			
4	3	68408.20	No	0.713578			
5	4	69787.50	No	0.129834			
6	5	72801.60	Yes	0.385910			
7	6	71767.70	No	0.530632			
8	7	78346.60	Yes	0.541793			
9	8	66670.20	No	0.808481			
10	9	70246.80	Yes	0.138579			
11	10	71255.00	No	0.492472			
12	11	72546.60	No	0.066401			
13	12	69512.50	Yes	0.847944			
14	13	71753.00	Yes	0.020783			
15	14	73547.10	No	0.627402			
16	15	68052.20	No	0.646576			
17	16	64652.50	Yes	0.816150			
18	17	71764.90	Yes	0.588482			
19	18	65187.80	Yes	0.668991			
20	19	69867.50	Yes	0.805540			
21	20	73706.30	Yes	0.391420			
22	21	72039.50	Yes	0.771441			
23	22	72973.60	No	0.998373			
24	23	73372.50	No	0.907789			
25	24	74592.00	Yes	0.059995			
26	25	75738.10	Yes	0.039803			
27	26	72975.10	Yes	0.093363			
28	27	72386.20	Yes	0.006328			
29	28	71051.60	Yes	0.859665			
30	29	72095.60	Yes	0.902875			
31	30	64956.50	No	0.392367			

The formula in cells D2:D2501 is =RAND()

	A	B	C	D	E
1	Employee	Annual Salary	Training Program	Random Numbers	
2	1721	64255.30	No	0.000404	
3	1293	80798.10	Yes	0.000957	
4	163	80859.10	Yes	0.003442	
5	733	65561.20	Yes	0.003520	
6	99	70788.80	Yes	0.003553	
7	652	69270.20	No	0.004620	
8	1709	75408.60	Yes	0.005811	
9	2071	73343.00	Yes	0.006122	
10	27	72386.20	Yes	0.006328	
11	239	67616.20	No	0.007030	
12	2092	70328.30	Yes	0.007629	
13	1054	74703.70	Yes	0.008555	
14	1004	72088.20	No	0.008943	
15	39	66940.10	Yes	0.009327	
16	1687	62442.20	No	0.009716	
17	1028	71641.80	No	0.009966	
18	400	73544.10	No	0.010106	
19	2318	70245.10	No	0.010160	
20	1403	74156.10	Yes	0.010624	
21	1228	64714.00	Yes	0.010645	
22	1434	68267.80	No	0.010850	
23	2213	73771.60	Yes	0.011290	
24	1368	61173.10	No	0.011625	
25	1474	69740.20	No	0.011766	
26	1247	67716.80	Yes	0.011918	
27	838	63406.50	Yes	0.013592	
28	1509	65742.70	No	0.013683	
29	1270	67485.90	Yes	0.013983	
30	813	71413.70	No	0.014455	
31	1141	73457.10	No	0.014879	

Note: Rows 32–2501 are not shown.

Care and judgment must be exercised in implementing the selection process for obtaining a random sample from an infinite population. Each case may require a different selection procedure. Let us consider two examples to see what we mean by the conditions: (1) Each element selected comes from the same population, and (2) each element is selected independently.

A common quality-control application involves a production process for which there is no limit on the number of elements that can be produced. The conceptual population from which we are sampling is all the elements that could be produced (not just the ones that are produced) by the ongoing production process. Because we cannot develop a list of all the elements that could be produced, the population is considered infinite. To be more specific, let us consider a production line designed to fill boxes with breakfast cereal to a mean weight of 24 ounces per box. Samples of 12 boxes filled by this process are periodically selected by a quality-control inspector to determine if the process is operating properly or whether, perhaps, a machine malfunction has caused the process to begin underfilling or overfilling the boxes.

With a production operation such as this, the biggest concern in selecting a random sample is to make sure that condition 1, the sampled elements are selected from the same population, is satisfied. To ensure that this condition is satisfied, the boxes must be selected at approximately the same point in time. This way the inspector avoids the possibility of selecting some boxes when the process is operating properly and other boxes when the process is not operating properly and is underfilling or overfilling the boxes. With a production process such as this, the second condition, each element is selected independently,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

is satisfied by designing the production process so that each box of cereal is filled independently. With this assumption, the quality-control inspector need only worry about satisfying the same population condition.

As another example of selecting a random sample from an infinite population, consider the population of customers arriving at a fast-food restaurant. Suppose an employee is asked to select and interview a sample of customers in order to develop a profile of customers who visit the restaurant. The customer-arrival process is ongoing, and there is no way to obtain a list of all customers in the population. So, for practical purposes, the population for this ongoing process is considered infinite. As long as a sampling procedure is designed so that all the elements in the sample are customers of the restaurant and they are selected independently, a random sample will be obtained. In this case, the employee collecting the sample needs to select the sample from people who come into the restaurant and make a purchase to ensure that the same population condition is satisfied. If, for instance, the person selected for the sample is someone who came into the restaurant just to use the restroom, that person would not be a customer and the same population condition would be violated. So, as long as the interviewer selects the sample from people making a purchase at the restaurant, condition 1 is satisfied. Ensuring that the customers are selected independently can be more difficult.

The purpose of the second condition of the random sample selection procedure (each element is selected independently) is to prevent selection bias. In this case, selection bias would occur if the interviewer were free to select customers for the sample arbitrarily. The interviewer might feel more comfortable selecting customers in a particular age group and might avoid customers in other age groups. Selection bias would also occur if the interviewer selected a group of five customers who entered the restaurant together and asked all of them to participate in the sample. Such a group of customers would be likely to exhibit similar characteristics, which might provide misleading information about the population of customers. Selection bias such as this can be avoided by ensuring that the selection of a particular customer does not influence the selection of any other customer. In other words, the elements (customers) are selected independently.

McDonald's, a fast-food restaurant chain, implemented a random sampling procedure for this situation. The sampling procedure was based on the fact that some customers presented discount coupons. Whenever a customer presented a discount coupon, the next customer served was asked to complete a customer profile questionnaire. Because arriving customers presented discount coupons randomly and independently of other customers, this sampling procedure ensured that customers were selected independently. As a result, the sample satisfied the requirements of a random sample from an infinite population.

Situations involving sampling from an infinite population are usually associated with a process that operates over time. Examples include parts being manufactured on a production line, repeated experimental trials in a laboratory, transactions occurring at a bank, telephone calls arriving at a technical support center, and customers entering a retail store. In each case, the situation may be viewed as a process that generates elements from an infinite population. As long as the sampled elements are selected from the same population and are selected independently, the sample is considered a random sample from an infinite population.

Notes + Comments

1. In this section we have been careful to define two types of samples: a simple random sample from a finite population and a random sample from an infinite population. In the remainder of the text, we will generally refer to both of these as either a *random sample* or simply a *sample*. We will not make a distinction of the sample being a "simple" random sample unless it is necessary for the exercise or discussion.
2. Statisticians who specialize in sample surveys from finite populations use sampling methods that provide probability samples. With a probability sample, each possible sample has a known probability of selection and a random process is used to select the elements for the sample. Simple random sampling is one of these methods. We use the term *simple* in simple random sampling to clarify that this is the probability sampling method that

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

ensures that each sample of size n has the same probability of being selected.

3. The number of different simple random samples of size n that can be selected from a finite population of size N is:

$$\frac{N!}{n!(N-n)!}$$

In this formula, $N!$ and $n!$ are the factorial formulas. For the EAI problem with $N = 2{,}500$ and $n = 30$, this expression can be used to show that approximately 2.75×10^{69} different simple random samples of 30 EAI employees can be obtained.

4. In addition to simple random sampling, other probability sampling methods include the following:
   - Stratified random sampling—a method in which the population is first divided into homogeneous subgroups or strata and then a simple random sample is taken from each stratum.
   - Cluster sampling—a method in which the population is first divided into heterogeneous subgroups or clusters and then simple random samples are taken from some or all of the clusters.
   - Systematic sampling—a method in which we sort the population based on an important characteristic, randomly select one of the first k elements of the population, and then select every kth element from the population thereafter.

   Calculation of sample statistics such as the sample mean $\bar{x}$, the sample standard deviation s, and the sample proportion $\bar{p}$ differ depending on which method of probability sampling is used. See specialized books on sampling such as *Elementary Survey Sampling* (2011) by Scheaffer, Mendenhall, and Ott for more information.
5. Nonprobability sampling methods include the following:
   - Convenience sampling—a method in which sample elements are selected on the basis of accessibility.
   - Judgment sampling—a method in which sample elements are selected based on the opinion of the person doing the study.

   Although nonprobability samples have the advantages of relatively easy sample selection and data collection, no statistically justified procedure allows a probability analysis or inference about the quality of nonprobability sample results. Statistical methods designed for probability samples should not be applied to a nonprobability sample, and we should be cautious in interpreting the results when a nonprobability sample is used to make inferences about a population.

7.2 Point Estimation

Now that we have described how to select a simple random sample, let us return to the EAI problem. A simple random sample of 30 employees and the corresponding data on annual salary and management training program participation are as shown in Table 7.1. The notation x_1, x_2, and so on is used to denote the annual salary of the first employee in the sample, the annual salary of the second employee in the sample, and so on. Participation in the management training program is indicated by Yes in the management training program column.

The mean and standard deviation of a sample are discussed in Chapter 2.

To estimate the value of a population parameter, we compute a corresponding characteristic of the sample, referred to as a **sample statistic**. For example, to estimate the population mean μ and the population standard deviation σ for the annual salary of EAI employees, we use the data in Table 7.1 to calculate the corresponding sample statistics: the sample mean and the sample standard deviation s. The sample mean is

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{2{,}154{,}420}{30} = \$71{,}814$$

and the sample standard deviation is

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{325{,}009{,}260}{29}} = \$3{,}384$$

To estimate p, the proportion of employees in the population who completed the management training program, we use the corresponding sample proportion $\bar{p}$. Let x denote the number of employees in the sample who completed the management training program. The data in Table 7.1 show that $x = 19$. Thus, with a sample size of $n = 30$, the sample proportion is

$$\bar{p} = \frac{x}{n} = \frac{19}{30} = 0.63$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 7.1 Annual Salary and Training Program Status for a Simple Random Sample of 30 EAI Employees

Annual Salary ($)	Management Training Program	Annual Salary ($)	Management Training Program
$x_1 = 69,094.30$	Yes	$x_{16} = 71,766.00$	Yes
$x_2 = 73,263.90$	Yes	$x_{17} = 72,541.30$	No
$x_3 = 69,343.50$	Yes	$x_{18} = 64,980.00$	Yes
$x_4 = 69,894.90$	Yes	$x_{19} = 71,932.60$	Yes
$x_5 = 67,621.60$	No	$x_{20} = 72,973.00$	Yes
$x_6 = 75,924.00$	Yes	$x_{21} = 65,120.90$	Yes
$x_7 = 69,092.30$	Yes	$x_{22} = 71,753.00$	Yes
$x_8 = 71,404.40$	Yes	$x_{23} = 74,391.80$	No
$x_9 = 70,957.70$	Yes	$x_{24} = 70,164.20$	No
$x_{10} = 75,109,70$	Yes	$x_{25} = 72,973.60$	No
$x_{11} = 65,922.60$	Yes	$x_{26} = 70,241.30$	No
$x_{12} = 77,268.40$	No	$x_{27} = 72,793.90$	No
$x_{13} = 75,688.80$	Yes	$x_{28} = 70,979.40$	Yes
$x_{14} = 71,564.70$	No	$x_{29} = 75,860.90$	Yes
$x_{15} = 76,188.20$	No	$x_{30} = 77,309.10$	No

By making the preceding computations, we perform the statistical procedure called *point estimation*. We refer to the sample mean $\bar{x}$ as the **point estimator** of the population mean μ, the sample standard deviation s as the point estimator of the population standard deviation σ, and the sample proportion $\bar{p}$ as the point estimator of the population proportion p. The numerical value obtained for $\bar{x}$, s, or $\bar{p}$ is called the **point estimate**. Thus, for the simple random sample of 30 EAI employees shown in Table 7.1, \$71,814 is the point estimate of μ, \$3,348 is the point estimate of σ, and 0.63 is the point estimate of p. Table 7.2 summarizes the sample results and compares the point estimates to the actual values of the population parameters.

Table 7.2 Summary of Point Estimates Obtained from a Simple Random Sample of 30 EAI Employees

Population Parameter	Parameter Value	Point Estimator	Point Estimate
μ = Population mean annual salary	\$71,800	$\bar{x}$ = Sample mean annual salary	\$71,814
σ = Population standard deviation for annual salary	\$4,000	s = Sample standard deviation for annual salary	\$3,348
p = Population proportion completing the management training program	0.60	$\bar{p}$ = Sample proportion having completed the management training program	0.63

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

As is evident from Table 7.2, the point estimates differ somewhat from the values of corresponding population parameters. This difference is to be expected because a sample, and not a census of the entire population, is being used to develop the point estimates.

Practical Advice

The subject matter of most of the rest of the book is concerned with statistical inference, of which point estimation is a form. We use a sample statistic to make an inference about a population parameter. When making inferences about a population based on a sample, it is important to have a close correspondence between the sampled population and the target population. The **target population** is the population about which we want to make inferences, while the sampled population is the population from which the sample is actually taken. In this section, we have described the process of drawing a simple random sample from the population of EAI employees and making point estimates of characteristics of that same population. So the sampled population and the target population are identical, which is the desired situation. But in other cases, it is not as easy to obtain a close correspondence between the sampled and target populations.

Consider the case of an amusement park selecting a sample of its customers to learn about characteristics such as age and time spent at the park. Suppose all the sample elements were selected on a day when park attendance was restricted to employees of a large company. Then the sampled population would be composed of employees of that company and members of their families. If the target population we wanted to make inferences about were typical park customers over a typical summer, then we might encounter a significant difference between the sampled population and the target population. In such a case, we would question the validity of the point estimates being made. Park management would be in the best position to know whether a sample taken on a particular day was likely to be representative of the target population.

In summary, whenever a sample is used to make inferences about a population, we should make sure that the study is designed so that the sampled population and the target population are in close agreement. Good judgment is a necessary ingredient of sound statistical practice.

7.3 Sampling Distributions

In the preceding section we said that the sample mean $\bar{x}$ is the point estimator of the population mean μ, and the sample proportion $\bar{p}$ is the point estimator of the population proportion p. For the simple random sample of 30 EAI employees shown in Table 7.1, the point estimate of μ is $\bar{x} = \$71{,}814$ and the point estimate of p is $\bar{p} = 0.63$. Suppose we select another simple random sample of 30 EAI employees and obtain the following point estimates:

$$\text{Sample mean: } \bar{x} = \$72{,}670$$
$$\text{Sample proportion: } \bar{p} = 0.70$$

Note that different values of $\bar{x}$ and $\bar{p}$ were obtained. Indeed, a second simple random sample of 30 EAI employees cannot be expected to provide the same point estimates as the first sample.

Now, suppose we repeat the process of selecting a simple random sample of 30 EAI employees over and over again, each time computing the values of $\bar{x}$ and $\bar{p}$. Table 7.3 contains a portion of the results obtained for 500 simple random samples, and Table 7.4 shows the frequency and relative frequency distributions for the 500 values of $\bar{x}$. Figure 7.2 shows the relative frequency histogram for the $\bar{x}$ values.

A **random variable** is a quantity whose values are not known with certainty. Because the sample mean $\bar{x}$ is a quantity whose values are not known with certainty, the sample mean $\bar{x}$ is a random variable. As a result, just like other random variables, $\bar{x}$ has a mean or expected value, a standard deviation, and a probability distribution. Because the various

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 7.3 Values of $\bar{x}$ and $\bar{p}$ from 500 Simple Random Samples of 30 EAI Employees

Sample Number	Sample Mean ($\bar{x}$)	Sample Proportion ($\bar{p}$)
1	71,814	0.63
2	72,670	0.70
3	71,780	0.67
4	71,588	0.53
.	.	.
.	.	.
.	.	.
500	71,752	0.50

Table 7.4 Frequency and Relative Frequency Distributions of $\bar{x}$ from 500 Simple Random Samples of 30 EAI Employees

Mean Annual Salary ($)	Frequency	Relative Frequency
69,500.00–69,999.99	2	0.004
70,000.00–70,499.99	16	0.032
70,500.00–70,999.99	52	0.104
71,000.00–71,499.99	101	0.202
71,500.00–71,999.99	133	0.266
72,000.00–72,499.99	110	0.220
72,500.00–72,999.99	54	0.108
73,000.00–73,499.99	26	0.052
73,500.00–73,999.99	6	0.012
Totals:	500	1.000

The concept of a random variable is discussed in Chapter 2. The properties of random variables and their relationship to probability concepts are discussed in Chapter 5.

possible values of $\bar{x}$ are the result of different simple random samples, the probability distribution of $\bar{x}$ is called the **sampling distribution** of $\bar{x}$. Knowledge of this sampling distribution and its properties will enable us to make probability statements about how close the sample mean $\bar{x}$ is to the population mean μ.

Let us return to Figure 7.2. We would need to enumerate every possible sample of 30 employees and compute each sample mean to completely determine the sampling distribution of $\bar{x}$. However, the histogram of 500 values of $\bar{x}$ gives an approximation of this sampling distribution. From the approximation we observe the bell-shaped appearance of the distribution. We note that the largest concentration of the $\bar{x}$ values and the mean of the 500 values of $\bar{x}$ is near the population mean $\mu = \$71{,}800$. We will describe the properties of the sampling distribution of $\bar{x}$ more fully in the next section.

The ability to understand the material in subsequent sections of this chapter depends heavily on the ability to understand and use the sampling distributions presented in this section.

The 500 values of the sample proportion $\bar{p}$ are summarized by the relative frequency histogram in Figure 7.3. As in the case of $\bar{x}$, $\bar{p}$ is a random variable. If every possible sample of size 30 were selected from the population and if a value of $\bar{p}$ were computed for each sample, the resulting probability distribution would be the sampling distribution of $\bar{p}$. The relative frequency histogram of the 500 sample values in Figure 7.3 provides a general idea of the appearance of the sampling distribution of $\bar{p}$.

In practice, we select only one simple random sample from the population. We repeated the sampling process 500 times in this section simply to illustrate that many different

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.2 Relative Frequency Histogram of $\bar{x}$ Values from 500 Simple Random Samples of Size 30 Each

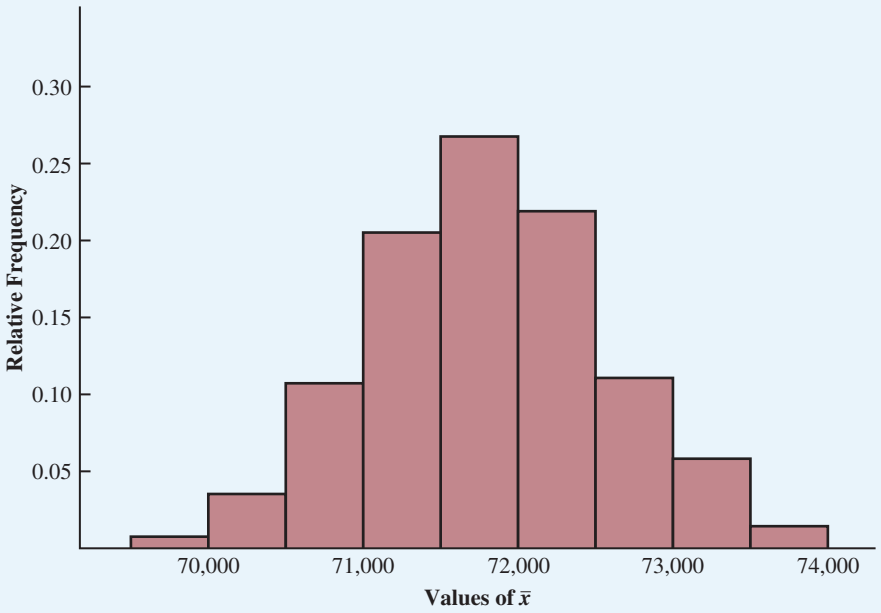
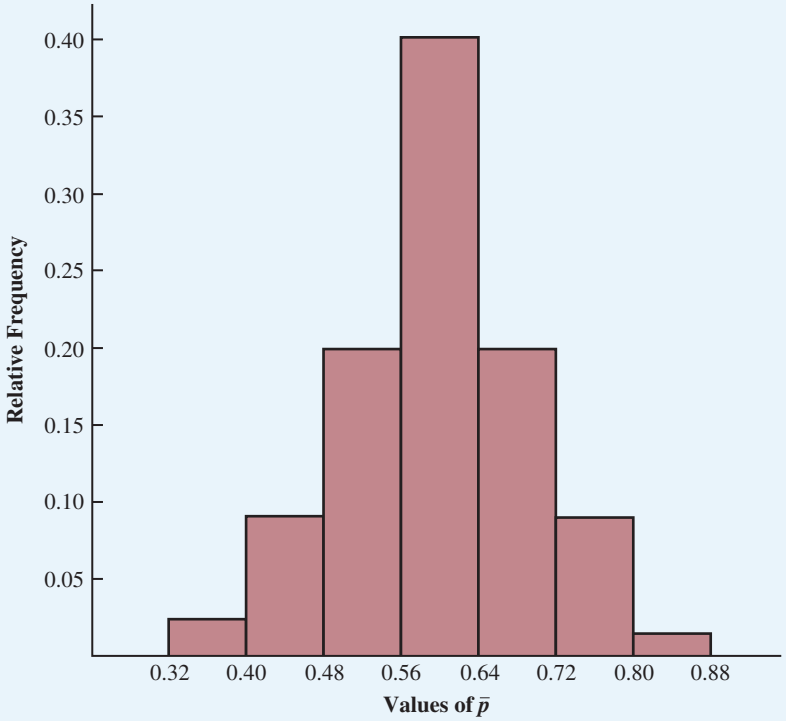


Figure 7.3 Relative Frequency Histogram of $\bar{p}$ Values from 500 Simple Random Samples of Size 30 Each



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

samples are possible and that the different samples generate a variety of values for the sample statistics $\bar{x}$ and $\bar{p}$. The probability distribution of any particular sample statistic is called the sampling distribution of the statistic. Next we discuss the characteristics of the sampling distributions of $\bar{x}$ and $\bar{p}$.

Sampling Distribution of $\bar{x}$

In the previous section we said that the sample mean $\bar{x}$ is a random variable and that its probability distribution is called the sampling distribution of $\bar{x}$.

Sampling Distribution of $\bar{x}$

The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$.

This section describes the properties of the sampling distribution of $\bar{x}$. Just as with other probability distributions we studied, the sampling distribution of $\bar{x}$ has an expected value or mean, a standard deviation, and a characteristic shape or form. Let us begin by considering the mean of all possible $\bar{x}$ values, which is referred to as the expected value of $\bar{x}$.

Expected Value of $\bar{x}$ In the EAI sampling problem we saw that different simple random samples result in a variety of values for the sample mean $\bar{x}$. Because many different values of the random variable $\bar{x}$ are possible, we are often interested in the mean of all possible values of $\bar{x}$ that can be generated by the various simple random samples. The mean of the $\bar{x}$ random variable is the expected value of $\bar{x}$. Let $E(\bar{x})$ represent the expected value of $\bar{x}$ and μ represent the mean of the population from which we are selecting a simple random sample. It can be shown that with simple random sampling, $E(\bar{x})$ and μ are equal.

The expected value of $\bar{x}$ equals the mean of the population from which the sample is selected.

Expected Value of $\bar{x}$

$$E(\bar{x}) = \mu \tag{7.1}$$

where

$E(\bar{x})$ = the expected value of $\bar{x}$
μ = the population mean

This result states that with simple random sampling, the expected value or mean of the sampling distribution of $\bar{x}$ is equal to the mean of the population. In Section 7.1 we saw that the mean annual salary for the population of EAI employees is $\mu = \$71{,}800$. Thus, according to equation (7.1), if we considered all possible samples of size n from the population of EAI employees, the mean of all the corresponding sample means for the EAI study would be equal to \$71,800, the population mean.

When the expected value of a point estimator equals the population parameter, we say the point estimator is **unbiased**. Thus, equation (7.1) states that $\bar{x}$ is an unbiased estimator of the population mean μ.

The term standard error is used in statistical inference to refer to the standard deviation of a point estimator.

Standard Deviation of $\bar{x}$ Let us define the standard deviation of the sampling distribution of $\bar{x}$. We will use the following notation:

$\sigma_{\bar{x}}$ = the standard deviation of $\bar{x}$, or the **standard error** of the mean
σ = the standard deviation of the population
n = the sample size
N = the population size

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

It can be shown that the formula for the standard deviation of $\bar{x}$ depends on whether the population is finite or infinite. The two formulas for the standard deviation of $\bar{x}$ follow.

Standard Deviation of $\bar{x}$

Finite Population	*Infinite Population*	
$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	**(7.2)**

In comparing the two formulas in equation (7.2), we see that the factor $\sqrt{(N-n)/(N-1)}$ is required for the finite population case but not for the infinite population case. This factor is commonly referred to as the **finite population correction factor**. In many practical sampling situations, we find that the population involved, although finite, is large relative to the sample size. In such cases the finite population correction factor $\sqrt{(N-n)/(N-1)}$ is close to 1. As a result, the difference between the values of the standard deviation of $\bar{x}$ for the finite and infinite populations becomes negligible. Then, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ becomes a good approximation to the standard deviation of $\bar{x}$ even though the population is finite. In cases where $n/N > 0.05$, the finite population version of equation (7.2) should be used in the computation of $\sigma_{\bar{x}}$. Unless otherwise noted, throughout the text we will assume that the population size is large relative to the sample size, i.e., $n/N \leq 0.05$.

Observe from equation (7.2) that we need to know σ, the standard deviation of the population, in order to compute $\sigma_{\bar{x}}$. That is, the sample-to-sample variability in the point estimator $\bar{x}$, as measured by the standard error $\sigma_{\bar{x}}$, depends on the standard deviation of the population from which the sample is drawn. However, when we are sampling to estimate the population mean with $\bar{x}$, usually the population standard deviation is also unknown. Therefore, we need to estimate the standard deviation of $\bar{x}$ with $s_{\bar{x}}$ using the sample standard deviations as shown in equation (7.3).

Estimated Standard Deviation of $\bar{x}$

Finite Population	*Infinite Population*	
$s_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{s}{\sqrt{n}}\right)$	$s_{\bar{x}} = \left(\frac{s}{\sqrt{n}}\right)$	**(7.3)**

Let us now return to the EAI example and compute the estimated standard error (standard deviation) of the mean associated with simple random samples of 30 EAI employees. Recall from Table 7.2 that the standard deviation of the sample of 30 EAI employees is $s = 3{,}348$. In this case, the population is finite ($N = 2{,}500$), but because $n/N = 30/2{,}500 = 0.012 < 0.05$, we can ignore the finite population correction factor and compute the estimated standard error as

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{3{,}348}{\sqrt{30}} = 611.3$$

In this case, we happen to know that the standard deviation of the population is actually $\sigma = 4{,}000$, so the true standard error is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4{,}000}{\sqrt{30}} = 730.3$$

The difference between $s_{\bar{x}}$ and $\sigma_{\bar{x}}$ is due to **sampling error**, or the error that results from observing a sample of 30 rather than the entire population of 2,500.

Form of the Sampling Distribution of $\bar{x}$ The preceding results concerning the expected value and standard deviation for the sampling distribution of $\bar{x}$ are applicable for any population. The final step in identifying the characteristics of the sampling distribution of $\bar{x}$ is

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to determine the form or shape of the sampling distribution. We will consider two cases: (1) The population has a normal distribution and (2) the population does not have a normal distribution.

Population Has a Normal Distribution In many situations it is reasonable to assume that the population from which we are selecting a random sample has a normal, or nearly normal, distribution. When the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed for any sample size.

Population Does Not Have a Normal Distribution When the population from which we are selecting a random sample does not have a normal distribution, the central limit theorem is helpful in identifying the shape of the sampling distribution of $\bar{x}$. A statement of the central limit theorem as it applies to the sampling distribution of x follows.

Central Limit Theorem

In selecting random samples of size n from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by a *normal distribution* as the sample size becomes large.

Figure 7.4 shows how the central limit theorem works for three different populations; each column refers to one of the populations. The top panel of the figure shows that none of the populations are normally distributed. Population I follows a uniform distribution. Population II is often called the rabbit-eared distribution. It is symmetric, but the more likely values fall in the tails of the distribution. Population III is shaped like the exponential distribution; it is skewed to the right.

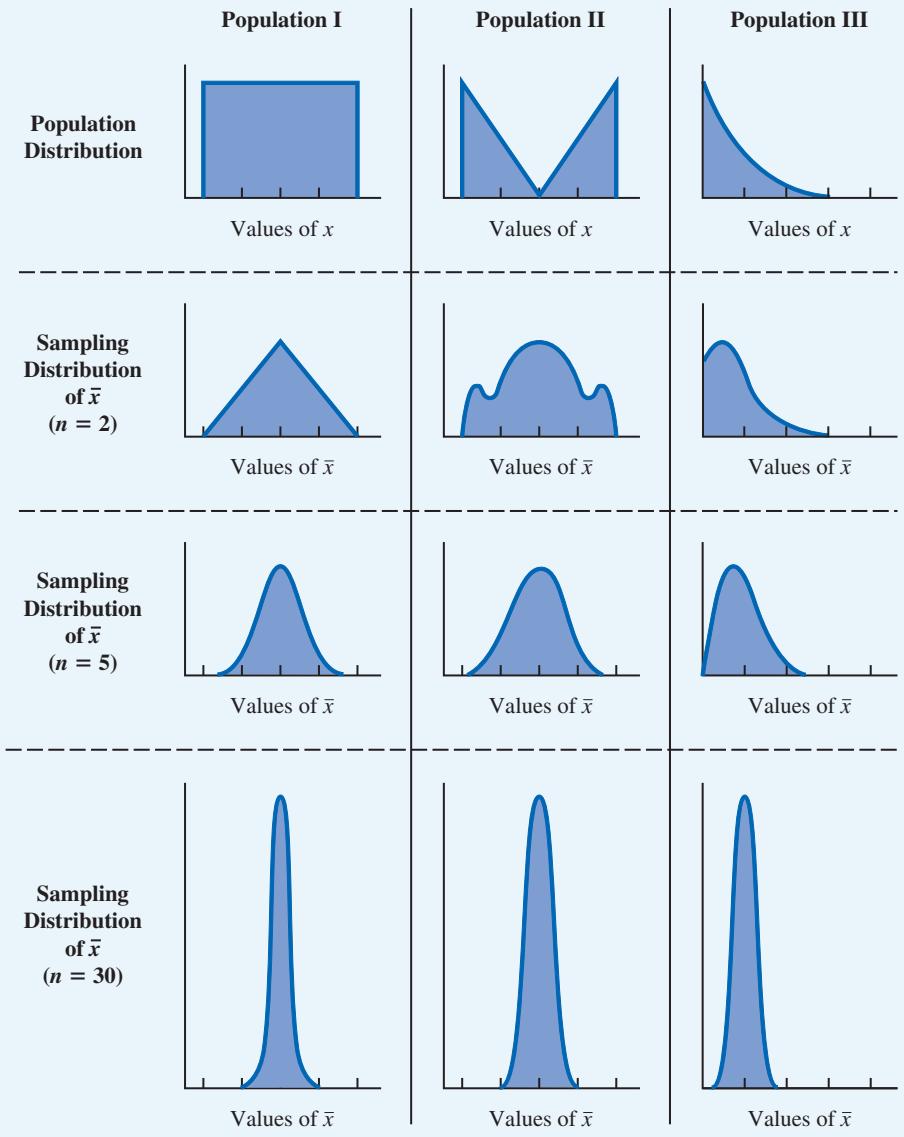
The bottom three panels of Figure 7.4 show the shape of the sampling distribution for samples of size $n = 2$, $n = 5$, and $n = 30$. When the sample size is 2, we see that the shape of each sampling distribution is different from the shape of the corresponding population distribution. For samples of size 5, we see that the shapes of the sampling distributions for populations I and II begin to look similar to the shape of a normal distribution. Even though the shape of the sampling distribution for population III begins to look similar to the shape of a normal distribution, some skewness to the right is still present. Finally, for a sample size of 30, the shapes of each of the three sampling distributions are approximately normal.

From a practitioner's standpoint, we often want to know how large the sample size needs to be before the central limit theorem applies and we can assume that the shape of the sampling distribution is approximately normal. Statistical researchers have investigated this question by studying the sampling distribution of $\bar{x}$ for a variety of populations and a variety of sample sizes. General statistical practice is to assume that, for most applications, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution whenever the sample size is 30 or more. In cases in which the population is highly skewed or outliers are present, sample sizes of 50 may be needed.

Sampling Distribution of $\bar{x}$ for the EAI Problem Let us return to the EAI problem where we previously showed that $E(\bar{x}) = \$71{,}800$ and $\sigma_{\bar{x}} = 730.3$. At this point, we do not have any information about the population distribution; it may or may not be normally distributed. If the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed. If the population does not have a normal distribution, the simple random sample of 30 employees and the central limit theorem enable us to conclude that the sampling distribution of $\bar{x}$ can be approximated by a normal distribution. In either case, we are comfortable proceeding with the conclusion that the sampling distribution of $\bar{x}$ can be described by the normal distribution shown in Figure 7.5. In other words, Figure 7.5 illustrates the distribution of the sample means corresponding to all possible sample sizes of 30 for the EAI study.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.4 Illustration of the Central Limit Theorem for Three Populations



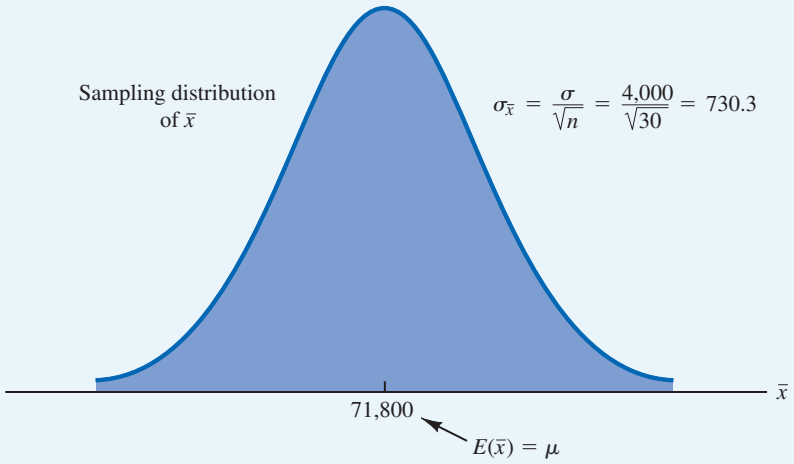
Relationship Between the Sample Size and the Sampling Distribution of $\bar{x}$ Suppose that in the EAI sampling problem we select a simple random sample of 100 EAI employees instead of the 30 originally considered. Intuitively, it would seem that because the larger sample size provides more data, the sample mean based on $n = 100$ would provide a better estimate of the population mean than the sample mean based on $n = 30$. To see how much better, let us consider the relationship between the sample size and the sampling distribution of $\bar{x}$.

First, note that $E(\bar{x}) = \mu$ regardless of the sample size. Thus, the mean of all possible values of $\bar{x}$ is equal to the population mean μ regardless of the sample size μ. However, note that the standard error of the mean, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, is related to the square root of the sample size. Whenever the sample size is increased, the standard error of the mean $\sigma_{\bar{x}}$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The sampling distribution in Figure 7.5 is a theoretical construct, as typically the population mean and the population standard deviation are not known. Instead, we must estimate these parameters with the sample mean and the sample standard deviation, respectively.

Figure 7.5 Sampling Distribution of $\bar{x}$ for the Mean Annual Salary of a Simple Random Sample of 30 EAI Employees

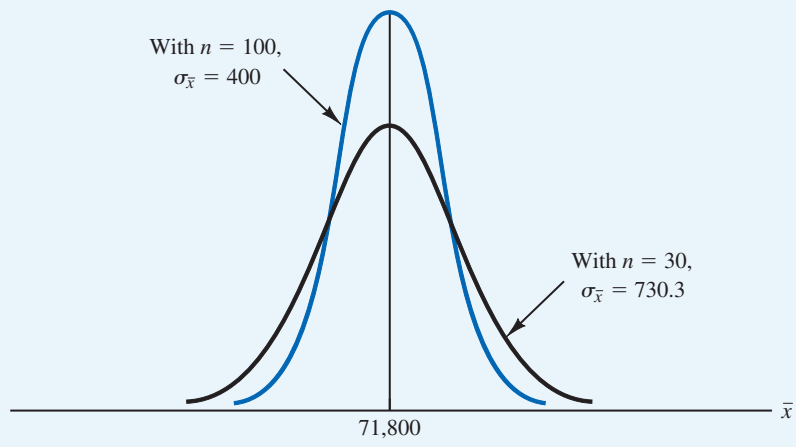


decreases. With $n = 30$, the standard error of the mean for the EAI problem is 730.3. However, with the increase in the sample size to $n = 100$, the standard error of the mean is decreased to

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4{,}000}{\sqrt{100}} = 400$$

The sampling distributions of $\bar{x}$ with $n = 30$ and $n = 100$ are shown in Figure 7.6. Because the sampling distribution with $n = 100$ has a smaller standard error, the values of $\bar{x}$ with $n = 100$ have less variation and tend to be closer to the population mean than the values of $\bar{x}$ with $n = 30$.

Figure 7.6 A Comparison of the Sampling Distributions of $\bar{x}$ for Simple Random Samples of $n = 30$ and $n = 100$ EAI Employees



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The important point in this discussion is that as the sample size increases, the standard error of the mean decreases. As a result, a larger sample size will provide a higher probability that the sample mean falls within a specified distance of the population mean. The practical reason we are interested in the sampling distribution of $\bar{x}$ is that it can be used to provide information about how close the sample mean is to the population mean. The concepts of interval estimation and hypothesis testing discussed in Sections 7.4 and 7.5 rely on the properties of sampling distributions.

Sampling Distribution of $\bar{p}$

The sample proportion $\bar{p}$ is the point estimator of the population proportion p. The formula for computing the sample proportion is

$$\bar{p} = \frac{x}{n}$$

where

x = the number of elements in the sample that possess the characteristic of interest
n = sample size

As previously noted in this section, the sample proportion $\bar{p}$ is a random variable and its probability distribution is called the sampling distribution of $\bar{p}$.

Sampling Distribution of $\bar{p}$

The sampling distribution of $\bar{p}$ is the probability distribution of all possible values of the sample proportion $\bar{p}$.

To determine how close the sample proportion $\bar{p}$ is to the population proportion p, we need to understand the properties of the sampling distribution of $\bar{p}$: the expected value of $\bar{p}$, the standard deviation of $\bar{p}$, and the shape or form of the sampling distribution of $\bar{p}$.

Expected Value of $\bar{p}$ The expected value of $\bar{p}$, the mean of all possible values of $\bar{p}$, is equal to the population proportion p.

Expected Value of $\bar{p}$

$$E(\bar{p}) = p \tag{7.4}$$

where

$E(\bar{p})$ = the expected value of $\bar{p}$
p = the population proportion

Because $E(\bar{p}) = p$, $\bar{p}$ is an unbiased estimator of p. In Section 7.1, we noted that $p = 0.60$ for the EAI population, where p is the proportion of the population of employees who participated in the company's management training program. Thus, the expected value of $\bar{p}$ for the EAI sampling problem is 0.60. That is, if we considered the sample proportions corresponding to all possible samples of size n for the EAI study, the mean of these sample proportions would be 0.6.

Standard Deviation of $\bar{p}$ Just as we found for the standard deviation of $\bar{x}$, the standard deviation of $\bar{p}$ depends on whether the population is finite or infinite. The two formulas for computing the standard deviation of $\bar{p}$ follow.

Standard Deviation of $\bar{p}$

Finite Population	*Infinite Population*
$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}$	$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$

(7.5)

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Comparing the two formulas in equation (7.5), we see that the only difference is the use of the finite population correction factor $\sqrt{(N-n)/(N-1)}$.

As was the case with the sample mean $\bar{x}$, the difference between the expressions for the finite population and the infinite population becomes negligible if the size of the finite population is large in comparison to the sample size. We follow the same rule of thumb that we recommended for the sample mean. That is, if the population is finite with $n/N \leq 0.05$, we will use $\sigma_{\bar{p}} = \sqrt{p(1-p)/n}$. However, if the population is finite with $n/N > 0.05$, the finite population correction factor should be used. Again, unless specifically noted, throughout the text we will assume that the population size is large in relation to the sample size and thus the finite population correction factor is unnecessary.

Earlier in this section, we used the term *standard error of the mean* to refer to the standard deviation of $\bar{x}$. We stated that in general the term *standard error* refers to the standard deviation of a point estimator. Thus, for proportions we use *standard error of the proportion* to refer to the standard deviation of $\bar{p}$. From equation (7.5), we observe that the sample-to-sample variability in the point estimator $\bar{p}$, as measured by the standard error $\sigma_{\bar{p}}$, depends on the population proportion p. However, when we are sampling to compute $\bar{p}$, typically the population proportion is unknown. Therefore, we need to estimate the standard deviation of $\bar{p}$ with $s_{\bar{p}}$ using the sample proportion as shown in equation (7.6).

Estimated Standard Deviation of $\bar{p}$

Finite Population	*Infinite Population*	
$s_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$	$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$	**(7.6)**

Let us now return to the EAI example and compute the estimated standard error (standard deviation) of the proportion associated with simple random samples of 30 EAI employees. Recall from Table 7.2 that the sample proportion of EAI employees who completed the management training program is $\bar{p} = 0.63$. Because $n/N = 30/2{,}500 = 0.012 < 0.05$, we can ignore the finite population correction factor and compute the estimated standard error as

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{0.63(1-0.63)}{30}} = 0.0881$$

In the EAI example, we actually know that the population proportion is $p = 0.6$, so we know that the true standard error is

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{30}} = 0.0894$$

The difference between $s_{\bar{p}}$ and $\sigma_{\bar{p}}$ is due to sampling error.

Form of the Sampling Distribution of $\bar{p}$ Now that we know the mean and standard deviation of the sampling distribution of $\bar{p}$, the final step is to determine the form or shape of the sampling distribution. The sample proportion is $\bar{p} = x/n$. For a simple random sample from a large population, x is a binomial random variable indicating the number of elements in the sample with the characteristic of interest. Because n is a constant, the probability of x/n is the same as the binomial probability of x, which means that the sampling distribution of $\bar{p}$ is also a discrete probability distribution and that the probability for each value of x/n the same as the binomial probability of the corresponding value of x.

Statisticians have shown that a binomial distribution can be approximated by a normal distribution whenever the sample size is large enough to satisfy the following two conditions:

$$np \geq 5 \quad \text{and} \quad n(1-p) \geq 5$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Because the population proportion p is typically unknown in a study, the test to see whether the sampling distribution of $\bar{p}$ can be approximated by a normal distribution is often based on the sample proportion, $n\bar{p} \geq 5$ and $n(1-\bar{p}) \geq 5$.

Assuming that these two conditions are satisfied, the probability distribution of x in the sample proportion, $\bar{p} = x/n$, can be approximated by a normal distribution. And because n is a constant, the sampling distribution of $\bar{p}$ can also be approximated by a normal distribution. This approximation is stated as follows:

The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever $np \geq 5$ and $n(1-p) \geq 5$.

In practical applications, when an estimate of a population proportion is desired, we find that sample sizes are almost always large enough to permit the use of a normal approximation for the sampling distribution of $\bar{p}$.

Recall that for the EAI sampling problem we know that a sample proportion of employees who participated in the training program is $\bar{p} = 0.63$. With a simple random sample of size 30, we have $n\bar{p} = 30(0.63) = 18.9$ and $n(1-\bar{p}) = 30(0.37) = 11.1$. Thus, the sampling distribution of $\bar{p}$ can be approximated by a normal distribution shown in Figure 7.7.

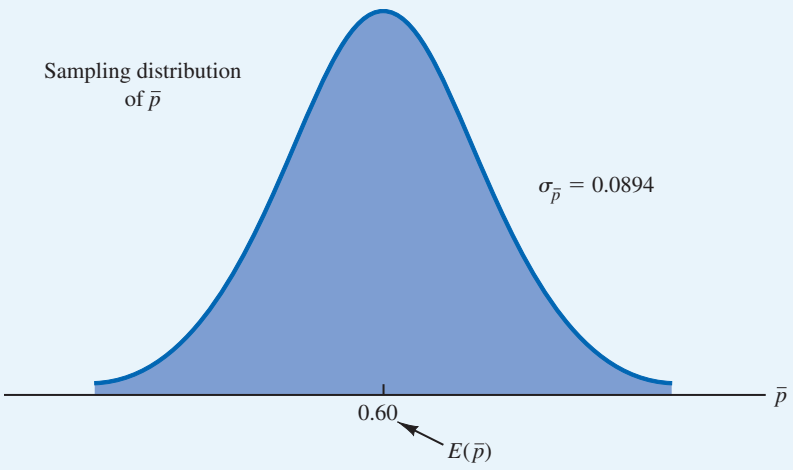
Relationship Between Sample Size and the Sampling Distribution of $\bar{p}$ Suppose that in the EAI sampling problem we select a simple random sample of 100 EAI employees instead of the 30 originally considered. Intuitively, it would seem that because the larger sample size provides more data, the sample proportion based on $n = 100$ would provide a better estimate of the population proportion than the sample proportion based on $n = 30$. To see how much better, recall that the standard error of the proportion is 0.0894 when the sample size is $n = 30$. If we increase the sample size to $n = 100$, the standard error of the proportion becomes

$$\sigma_{\bar{p}} = \sqrt{\frac{0.60(1-0.60)}{100}} = 0.0490$$

As we observed with the standard deviation of the sampling distribution of $\bar{x}$, increasing the sample size decreases the sample-to-sample variability of the sample proportion. As a result, a larger sample size will provide a higher probability that the sample proportion falls within a specified distance of the population proportion. The practical reason we are

The sampling distribution in Figure 7.7 is a theoretical construct, as typically the population proportion is not known. Instead, we must estimate it with the sample proportion.

Figure 7.7 Sampling Distribution of $\bar{p}$ for the Proportion of EAI Employees Who Participated in the Management Training Program



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

interested in the sampling distribution of $\overline{p}$ is that it can be used to provide information about how close the sample proportion is to the population proportion. The concepts of interval estimation and hypothesis testing discussed in Sections 7.4 and 7.5 rely on the properties of sampling distributions.

7.4 Interval Estimation

In Section 7.2, we stated that a point estimator is a sample statistic used to estimate a population parameter. For instance, the sample mean $\overline{x}$ is a point estimator of the population mean μ and the sample proportion $\overline{p}$ is a point estimator of the population proportion p. Because a point estimator cannot be expected to provide the exact value of the population parameter, **interval estimation** is frequently used to generate an estimate of the value of a population parameter. An **interval estimate** is often computed by adding and subtracting a value, called the **margin of error**, to the point estimate:

$$\text{Point estimate} \pm \text{Margin of error}$$

The purpose of an interval estimate is to provide information about how close the point estimate, provided by the sample, is to the value of the population parameter. In this section, we show how to compute interval estimates of a population mean μ and a population proportion p.

Interval Estimation of the Population Mean

The general form of an interval estimate of a population mean is

$$\overline{x} \pm \text{Margin of error}$$

The sampling distribution of $\overline{x}$ plays a key role in computing this interval estimate.

In Section 7.3 we showed that the sampling distribution of $\overline{x}$ has a mean equal to the population mean ($E(\overline{x}) = \mu$) and a standard deviation equal to the population standard deviation divided by the square root of the sample size ($\sigma_{\overline{x}} = \sigma/\sqrt{n}$). We also showed that for a sufficiently large sample or for a sample taken from a normally distributed population, the sampling distribution of $\overline{x}$ follows a normal distribution. These results for samples of 30 EAI employees are illustrated in Figure 7.5. Because the sampling distribution of $\overline{x}$ shows how values of $\overline{x}$ are distributed around the population mean μ, the sampling distribution of $\overline{x}$ provides information about the possible differences between $\overline{x}$ and μ.

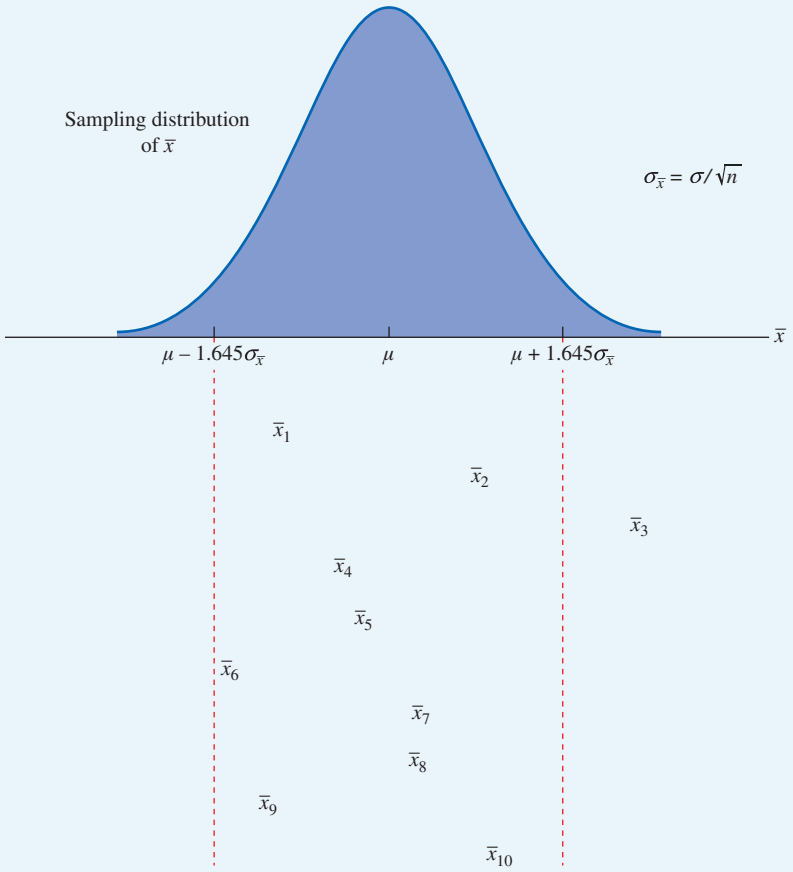
For any normally distributed random variable, 90% of the values lie within 1.645 standard deviations of the mean, 95% of the values lie within 1.960 standard deviations of the mean, and 99% of the values lie within 2.576 standard deviations of the mean. Thus, when the sampling distribution of $\overline{x}$ is normal, 90% of all values of $\overline{x}$ must be within $\pm 1.645\sigma_{\overline{x}}$ of the mean μ, 95% of all values of $\overline{x}$ must be within $\pm 1.96\sigma_{\overline{x}}$ of the mean μ, and 99% of all values of $\overline{x}$ must be within $\pm 2.576\sigma_{\overline{x}}$ of the mean μ.

Figure 7.8 shows what we would expect for values of sample means for 10 independent random samples when the sampling distribution of $\overline{x}$ is normal. Because 90% of all values of $\overline{x}$ are within $\pm 1.645\sigma_{\overline{x}}$ of the mean μ, we expect 9 of the values of $\overline{x}$ for these 10 samples to be within $\pm 1.645\sigma_{\overline{x}}$ of the mean μ. If we repeat this process of collecting 10 samples, our results may not include 9 sample means with values that are within $1.645\sigma_{\overline{x}}$ of the mean μ, but on average, the values of $\overline{x}$ will be within $\pm 1.645\sigma_{\overline{x}}$ of the mean μ for 9 of every 10 samples.

We now want to use what we know about the sampling distribution of $\overline{x}$ to develop an interval estimate of the population mean μ. However, when developing an interval estimate of a population mean μ, we generally do not know the population standard deviation σ, and therefore, we do not know the standard error of $\overline{x}$, $\sigma_{\overline{x}} = \sigma/\sqrt{n}$. In this case, we must use the same sample data to estimate both μ and σ, so we use $s_{\overline{x}} = s/\sqrt{n}$ to estimate the standard error of $\overline{x}$. When we estimate $\sigma_{\overline{x}}$ with $s_{\overline{x}}$, we introduce an additional source of uncertainty about the distribution of values of $\overline{x}$. If the sampling distribution of $\overline{x}$ follows a

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.8 Sampling Distribution of the Sample Mean



normal distribution, we address this additional source of uncertainty by using a probability distribution known as the ***t* distribution**.

The *t* distribution is a family of similar probability distributions; the shape of each specific *t* distribution depends on a parameter referred to as the **degrees of freedom**. The *t* distribution with 1 degree of freedom is unique, as is the *t* distribution with 2 degrees of freedom, the *t* distribution with 3 degrees of freedom, and so on. These *t* distributions are similar in shape to the **standard normal distribution** but are wider; this reflects the additional uncertainty that results from using $s_{\bar{x}}$ to estimate $\sigma_{\bar{x}}$. As the degrees of freedom increase, the difference between $s_{\bar{x}}$ and $\sigma_{\bar{x}}$ decreases and the *t* distribution narrows. Furthermore, because the area under any distribution curve is fixed at 1.0, a narrower *t* distribution will have a higher peak. Thus, as the degrees of freedom increase, the *t* distribution narrows, its peak becomes higher, and it becomes more similar to the standard normal distribution. We can see this in Figure 7.9, which shows *t* distributions with 10 and 20 degrees of freedom as well as the standard normal probability distribution. Note that as with the standard normal distribution, the mean of the *t* distribution is zero.

The standard normal distribution is a normal distribution with a mean of zero and a standard deviation of one. Chapter 6 contains a discussion of the normal distribution and the special case of the standard normal distribution.

To use the *t* distribution to compute the margin of error for the EAI example, we consider the *t* distribution with $n - 1 = 30 - 1 = 29$ degrees of freedom. Figure 7.10 shows that for a *t*-distributed random variable with 29 degrees of freedom, 90% of the values are within ±1.699 standard deviations of the mean and 10% of the values are more than ±1.699 standard deviations away from the mean. Thus, 5% of the values are more than 1.699 standard deviations below the mean and 5% of the values are more than 1.699

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Although the mathematical development of the t *distribution is based on the assumption that the population from which we are sampling is normally distributed, research shows that the* t *distribution can be successfully applied in many situations in which the population deviates substantially from a normal distribution.*

Figure 7.9 Comparison of the Standard Normal Distribution with *t* Distributions with 10 and 20 Degrees of Freedom

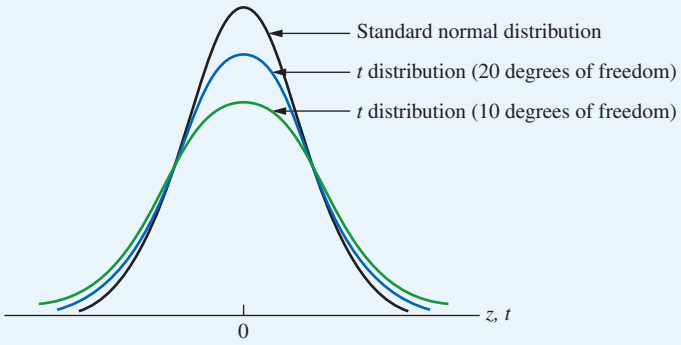
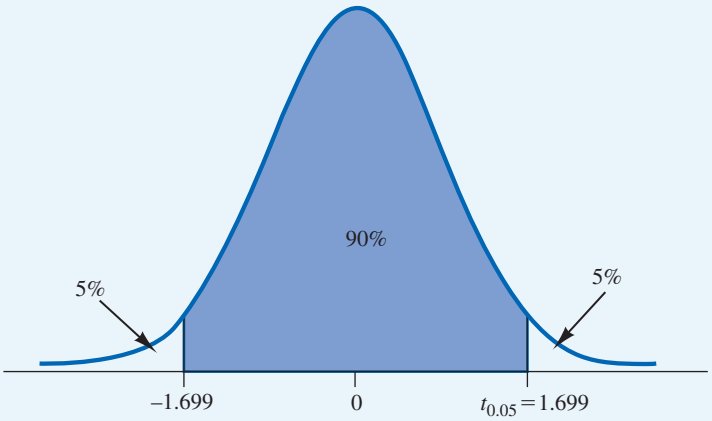


Figure 7.10 *t* Distribution with 29 Degrees of Freedom



standard deviations above the mean. This leads us to use $t_{0.05}$ to denote the value of t for which the area in the upper tail of a t distribution is 0.05. For a t distribution with 29 degrees of freedom, $t_{0.05} = 1.699$.

To see how the difference between the t *distribution and the standard normal distribution decreases as the degrees of freedom increase, use Excel's* T.INV.2T *function to compute* $t_{0.05}$ *for increasingly larger degrees of freedom* (n − 1) *and watch the value of* $t_{0.05}$ *approach 1.645.*

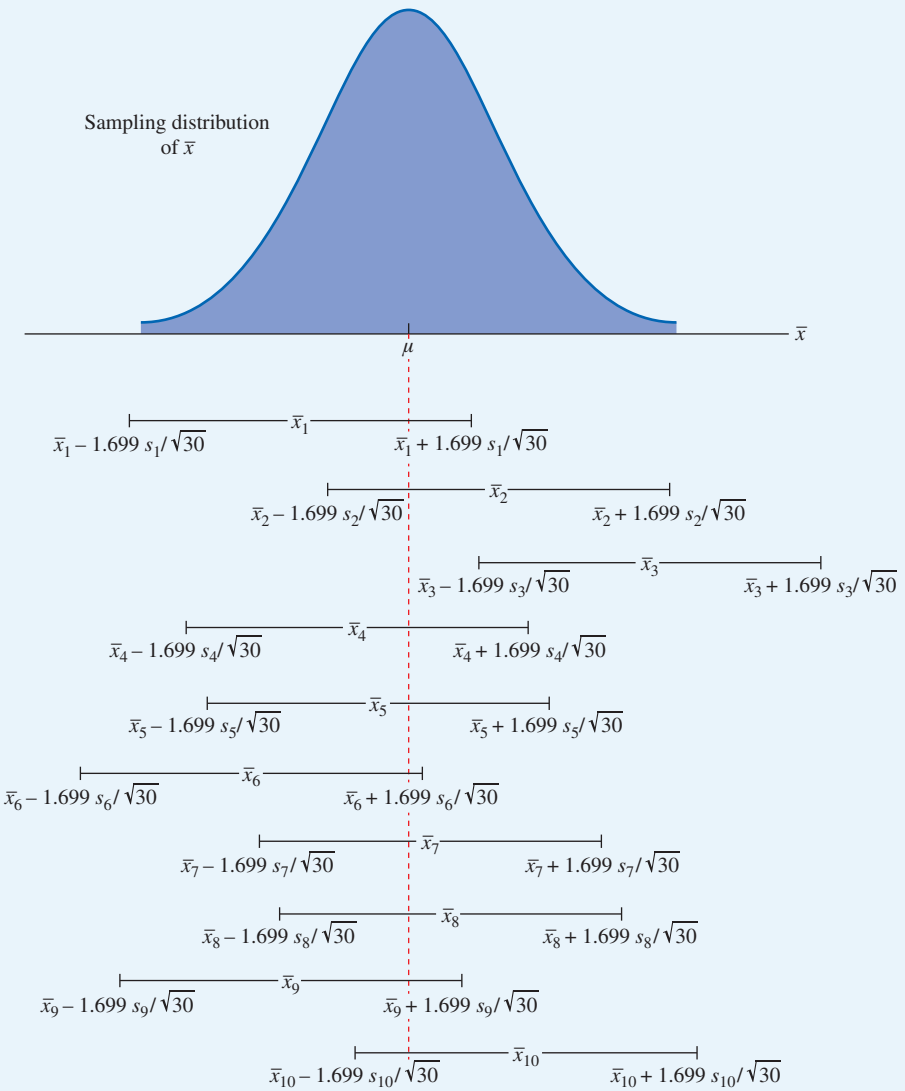
We can use Excel's T.INV.2T function to find the value from a t distribution such that a given percentage of the distribution is included in the interval $\pm t$ for any degrees of freedom. For example, suppose again that we want to find the value of t from the t distribution with 29 degrees of freedom such that 90% of the t distribution is included in the interval $-t$ to $+t$. Excel's T.INV.2T function has two inputs: (1) 1 − the proportion of the t distribution that will fall between $-t$ and $+t$, and (2) the degrees of freedom (which in this case is equal to the sample size − 1). For our example, we would enter the formula = *T.INV.2T(1 - 0.90, 30 - 1)*, which computes the value of 1.699. This confirms the data shown in Figure 7.10; for the t distribution with 29 degrees of freedom, $t_{0.05} = 1.699$ and 90% of all values for the t distribution with 29 degrees of freedom will lie between −1.699 and 1.699.

At the beginning of this section, we stated that the general form of an interval estimate of the population mean μ is $\bar{x} \pm$ margin of error. To provide an interpretation for this interval

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

estimate, let us consider the values of $\bar{x}$ that might be obtained if we took 10 independent simple random samples of 30 EAI employees. The first sample might have the mean $\bar{x}_1$ and standard deviation s_1. Figure 7.11 shows that the interval formed by subtracting $1.699s_1/\sqrt{30}$ from $\bar{x}_1$ and adding $1.699s_1/\sqrt{30}$ to $\bar{x}_1$ includes the population mean μ. Now consider what happens if the second sample has the mean $\bar{x}_2$ and standard deviation s_2. Although this sample mean differs from the first sample mean, we see in Figure 7.11 that the interval formed by subtracting $1.699s_2/\sqrt{30}$ from $\bar{x}_2$ and adding $1.699s_2/\sqrt{30}$ to $\bar{x}_2$ also includes the population mean μ. However, consider the third sample, which has the mean $\bar{x}_3$ and standard deviation s_3. As we see in Figure 7.11, the interval formed by subtracting $1.699s_3/\sqrt{30}$ from $\bar{x}_3$ and adding $1.699s_3/\sqrt{30}$ to $\bar{x}_3$ does not include the population mean μ. Because we are using $t_{0.05} = 1.699$ to form this interval, we expect that 90% of the intervals for our samples will include the population mean μ, and we see in Figure 7.11 that the results for our 10 samples of 30 EAI employees are what we would expect; the intervals for 9 of the 10 samples of $n = 30$ observations in this example include the mean μ. However, it is important to note that if we repeat

Figure 7.11 Intervals Formed Around Sample Means from 10 Independent Random Samples



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

this process of collecting 10 samples of $n = 30$ EAI employees, we may find that fewer than 9 of the resulting intervals $\bar{x} \pm 1.699 s_{\bar{x}}$ include the mean μ or all 10 of the resulting intervals $\bar{x} \pm 1.699 s_{\bar{x}}$ include the mean μ. However, on average, the resulting intervals $\bar{x} \pm 1.699 s_{\bar{x}}$ for 9 of 10 samples of $n = 30$ observations will include the mean μ.

Now recall that the sample of $n = 30$ EAI employees from Section 7.2 had a sample mean of salary of $\bar{x} = \$71{,}814$ and sample standard deviation of $s = \$3{,}340$. Using $\bar{x} \pm 1.699(3{,}340/\sqrt{30})$ to construct the interval estimate, we obtain $71{,}814 \pm 1{,}036$. Thus, the specific interval estimate of μ based on this specific sample is \$70,778 to \$72,850. Because approximately 90% of all the intervals constructed using $\bar{x} \pm 1.699(s/\sqrt{30})$ will contain the population mean, we say that we are approximately 90% confident that the interval \$70,778 to \$72,850 includes the population mean μ. We also say that this interval has been established at the 90% **confidence level**. The value of 0.90 is referred to as the **confidence coefficient**, and the interval \$70,564 to \$73,064 is called the 90% **confidence interval**.

Another term sometimes associated with an interval estimate is the **level of significance**. The level of significance associated with an interval estimate is denoted by the Greek letter α. The level of significance and the confidence coefficient are related as follows:

$$\alpha = \text{level of significance} = 1 - \text{confidence coefficient}$$

The level of significance is the probability that the interval estimation procedure will generate an interval that does not contain μ (such as the third sample in Figure 7.11). For example, the level of significance corresponding to a 0.90 confidence coefficient is $\alpha = 1 - 0.90 = 0.10$.

In general, we use the notation $t_{\alpha/2}$ to represent the value such that there is an area of $\alpha/2$ in the upper tail of the t distribution (see Figure 7.12). If the sampling distribution of $\bar{x}$ is normal, the margin of error for an interval estimate of a population mean μ is

$$t_{\alpha/2} s_{\bar{x}} = t_{\alpha/2} \frac{s}{\sqrt{n}}$$

So if the sampling distribution of $\bar{x}$ is normal, we find the interval estimate of the mean μ by subtracting this margin of error from the sample mean $\bar{x}$ and adding this margin of error to the sample mean $\bar{x}$. Using the notation we have developed, equation (7.7) can be used to find the confidence interval or interval estimate of the population mean μ.

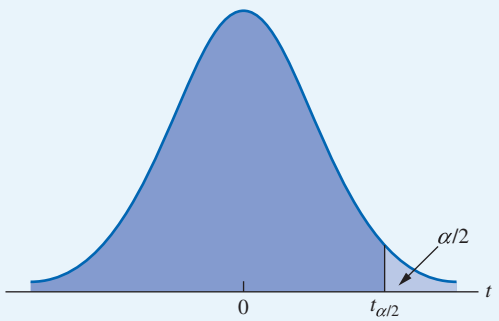
Observe that the margin of error, $t_{\alpha/2}(s/\sqrt{n})$, varies from sample to sample. This variation occurs because the sample standard deviation s varies depending on the sample selected. A large value for s results in a larger margin of error, while a small value for s results in a smaller margin of error.

Interval Estimate of a Population Mean

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}, \tag{7.7}$$

where s is the sample standard deviation, α is the level of significance, and $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of the t distribution with $n - 1$ degrees of freedom.

Figure 7.12 t Distribution with $\alpha/2$ Area or Probability in the Upper Tail



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If we want to find a 95% confidence interval for the mean μ in the EAI example, we again recognize that the degrees of freedom are $30 - 1 = 29$ and then use Excel's T.INV.2T function to find $t_{0.025} = 2.045$. We have seen that $s_{\bar{x}} = 611.3$ in the EAI example, so the margin of error at the 95% level of confidence is $t_{0.025}s_{\bar{x}} = \pm 2.045(611.3) = 1{,}250$. We also know that $\bar{x} = 71{,}814$ for the EAI example, so the 95% confidence interval is $71{,}814 \pm 1{,}250$, or \$70,564 to \$73,064.

It is important to note that a 95% confidence interval does not have a 95% probability of containing the population mean μ. Once constructed, a confidence interval will either contain the population parameter (μ in this EAI example) or not contain the population parameter. If we take several independent samples of the same size from our population and construct a 95% confidence interval for each of these samples, we would expect 95% of these confidence intervals to contain the mean μ. Our 95% confidence interval for the EAI example, \$70,564 to \$73,064, does indeed contain the population mean \$71,800; however, if we took many independent samples of 30 EAI employees and developed a 95% confidence interval for each, we would expect that 5% of these confidence intervals would not include the population mean \$71,800.

To further illustrate the interval estimation procedure, we will consider a study designed to estimate the mean credit card debt for the population of U.S. households. A sample of $n = 70$ households provided the credit card balances shown in Table 7.5. For this situation, no previous estimate of the population standard deviation σ is available. Thus, the sample data must be used to estimate both the population mean and the population standard deviation. Using the data in Table 7.5, we compute the sample mean $\bar{x} = \$9{,}312$ and the sample standard deviation $s = \$4{,}007$.

We can use Excel's T.INV.2T function to compute the value of $t_{\alpha/2}$ to use in finding this confidence interval. With a 95% confidence level and $n - 1 = 69$ degrees of freedom, we have that T.INV.2T$(1 - 0.95, 69) = 1.995$, so $t_{\alpha/2} = t_{(1-0.95)/2} = t_{0.025} = 1.995$ for this confidence interval.

We use equation (7.7) to compute an interval estimate of the population mean credit card balance.

$$9{,}312 \pm 1.995\frac{4{,}007}{\sqrt{70}}$$

$$9{,}312 \pm 995$$

The point estimate of the population mean is \$9,312, the margin of error is \$955, and the 95% confidence interval is $9{,}312 - 955 = \$8{,}357$ to $9{,}312 + 955 = \$10{,}267$. Thus, we are 95% confident that the mean credit card balance for the population of all households is between \$8,357 and \$10,267.

Using Excel We will use the credit card balances in Table 7.5 to illustrate how Excel can be used to construct an interval estimate of the population mean. We start by summarizing

DATA*file* newbalance

Table 7.5 Credit Card Balances for a Sample of 70 Households

9,430	14,661	7,159	9,071	9,691	11,032
7,535	12,195	8,137	3,603	11,448	6,525
4,078	10,544	9,467	16,804	8,279	5,239
5,604	13,659	12,595	13,479	5,649	6,195
5,179	7,061	7,917	14,044	11,298	12,584
4,416	6,245	11,346	6,817	4,353	15,415
10,676	13,021	12,806	6,845	3,467	15,917
1,627	9,719	4,972	10,493	6,191	12,591
10,112	2,200	11,356	615	12,851	9,743
6,567	10,746	7,117	13,627	5,337	10,324
13,627	12,744	9,465	12,557	8,372	
18,719	5,742	19,263	6,232	7,445	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the data using Excel's Descriptive Statistics tool. Refer to Figure 7.13 as we describe the tasks involved. The formula worksheet is on the left; the value worksheet is on the right.

If you can't find **Data Analysis** *on the* **Data** *tab, you may need to install the* **Analysis Toolpak** *add-in (which is included with Excel).*

Step 1. Click the **Data** tab on the Ribbon
Step 2. In the **Analysis** group, click **Data Analysis**
Step 3. When the **Data Analysis** dialog box appears, choose **Descriptive Statistics** from the list of Analysis Tools
Step 4. When the **Descriptive Statistics** dialog box appears:
Enter *A1:A71* in the **Input Range** box
Select **Grouped By Columns**
Select **Labels in First Row**
Select **Output Range:**
Enter *C1* in the **Output Range** box
Select **Summary Statistics**
Select **Confidence Level for Mean**
Enter *95* in the **Confidence Level for Mean** box
Click **OK**

Figure 7.13 95% Confidence Interval for Credit Card Balances

	A	B	C	D	E	F
1	**NewBalance**		*NewBalance*			
2	9430				Point Estimate	
3	7535		Mean	9312		
4	4078		Standard Error	478.928085819952		
5	5604		Median	9466		
6	5179		Mode	13627		
7	4416		Standard Deviation	4006.99984990035		
8	10676		Sample Variance	16056047.7971014		
9	1627		Kurtosis	-0.296002497492274		
10	10112		Skewness	0.187919587444176		
11	6567		Range	18648		
12	13627		Minimum	615		
13	18719		Maximum	19263		
14	14661		Sum	651840		
15	12195		Count	70		
16	10544		Confidence Level(95.0%)	955.435388972598	Margin of Error	
17	13659					
18	7061		**Point Estimate**	=D3		
19	6245		**Lower Limit**	=D18-D16		
20	13021		**Upper Limit**	=D3+D16		
70	9743					
71	10324					

	A	B	C	D	E
1	**NewBalance**		*NewBalance*		
2	9430				
3	7535		Mean	9312	
4	4078		Standard Error	478.9281	
5	5604		Median	9466	
6	5179		Mode	13627	
7	4416		Standard Deviation	4007	
8	10676		Sample Variance	16056048	
9	1627		Kurtosis	-0.296	
10	10112		Skewness	0.18792	
11	6567		Range	18648	
12	13627		Minimum	615	
13	18719		Maximum	19263	
14	14661		Sum	651840	
15	12195		Count	70	
16	10544		Confidence Level(95.0%)	955.4354	
17	13659				
18	7061		**Point Estimate**	9312	
19	6245		**Lower Limit**	8356.565	
20	13021		**Upper Limit**	10267.44	
70	9743				
71	10324				
72					

Note: Rows 21–69 are hidden.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The margin of error using the t *distribution can also be computed with the Excel function* CONFIDENCE.T(alpha, s, n), *where* alpha *is the level of significance,* s *is the sample standard deviation, and* n *is the sample size.*

As Figure 7.13 illustrates, the sample mean ($\bar{x}$) is in cell D3. The margin of error, labeled "Confidence Level(95%)," appears in cell D16. The value worksheet shows $\bar{x} = 9{,}312$ and a margin of error equal to 955.

Cells D18:D20 provide the point estimate and the lower and upper limits for the confidence interval. Because the point estimate is just the sample mean, the formula *=D3* is entered into cell D18. To compute the lower limit of the 95% confidence interval, $\bar{x}$ − (margin of error), we enter the formula *=D18-D16* into cell D19. To compute the upper limit of the 95% confidence interval, $\bar{x}$ + (margin of error), we enter the formula *=D18+D16* into cell D20. The value worksheet shows a lower limit of 8,357 and an upper limit of 10,267. In other words, the 95% confidence interval for the population mean is from 8,357 to 10,267.

Interval Estimation of the Population Proportion

The general form of an interval estimate of a population proportion p is

$$\bar{p} \pm \text{Margin of error}$$

The notation $z_{\alpha/2}$ *represents the value such that there is an area of* $\alpha/2$ *in the upper tail of the standard normal distribution (a normal distribution with a mean of zero and standard deviation of one).*

The sampling distribution of $\bar{p}$ plays a key role in computing the margin of error for this interval estimate.

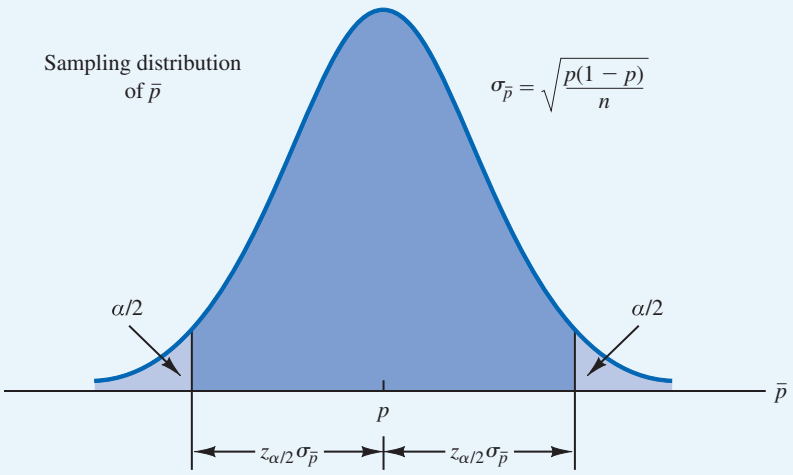
In Section 7.3 we said that the sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever $np \geq 5$ and $n(1-p) \geq 5$. Figure 7.14 shows the normal approximation of the sampling distribution of $\bar{p}$. The mean of the sampling distribution of $\bar{p}$ is the population proportion p, and the standard error of $\bar{p}$ is

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{7.8}$$

Because the sampling distribution of $\bar{p}$ is normally distributed, if we choose $z_{\alpha/2}\sigma_{\bar{p}}$ as the margin of error in an interval estimate of a population proportion, we know that $100(1-\alpha)\%$ of the intervals generated will contain the true population proportion. But $\sigma_{\bar{p}}$ cannot be used directly in the computation of the margin of error because p will not be known; p is what we are trying to estimate. So we estimate $\sigma_{\bar{p}}$ with $s_{\bar{p}}$ and then the margin of error for an interval estimate of a population proportion is given by

$$\text{Margin of error} = z_{\alpha/2}s_{\bar{p}} = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{7.9}$$

Figure 7.14 Normal Approximation of the Sampling Distribution of $\bar{p}$



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

With this margin of error, the general expression for an interval estimate of a population proportion is as follows.

Interval Estimate of a Population Proportion

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \tag{7.10}$$

where α is the level of significance and $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal distribution.

DATA*file* teetimes

The following example illustrates the computation of the margin of error and interval estimate for a population proportion. A national survey of 900 women golfers was conducted to learn how women golfers view their treatment at golf courses in the United States. The survey found that 396 of the women golfers were satisfied with the availability of tee times. Thus, the point estimate of the proportion of the population of women golfers who are satisfied with the availability of tee times is $396/900 = 0.44$. Using equation (7.10) and a 95% confidence level:

The Excel formula =NORM.S.INV(1 − α/2) *computes the value of* $z_{\alpha/2}$. *For example, for* $\alpha = 0.05$, $z_{0.0251}$ = NORM.S.INV (1 − .05/2) = 1.96.

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$0.44 \pm 1.96\sqrt{\frac{0.44(1-0.44)}{900}}$$

$$0.44 \pm 0.0324$$

Thus, the margin of error is 0.0324 and the 95% confidence interval estimate of the population proportion is 0.4076 to 0.4724. Using percentages, the survey results enable us to state with 95% confidence that between 40.76% and 47.24% of all women golfers are satisfied with the availability of tee times.

Using Excel Excel can be used to construct an interval estimate of the population proportion of women golfers who are satisfied with the availability of tee times. The responses in the survey were recorded as a Yes or No in the file *teetimes* for each woman surveyed. Refer to Figure 7.15 as we describe the tasks involved in constructing a 95% confidence interval. The formula worksheet is on the left; the value worksheet appears on the right.

The file teetimes *displayed in Figure 7.15 can be used as a template for developing confidence intervals about a population proportion* p, *by entering new problem data in column A and appropriately adjusting the formulas in column D.*

The descriptive statistics we need and the response of interest are provided in cells D3:D6. Because Excel's COUNT function works only with numerical data, we used the COUNTA function in cell D3 to compute the sample size. The response for which we want to develop an interval estimate, *Yes* or *No*, is entered into cell D4. Figure 7.15 shows that *Yes* has been entered into cell D4, indicating that we want to develop an interval estimate of the population proportion of women golfers who are satisfied with the availability of tee times. If we had wanted to develop an interval estimate of the population proportion of women golfers who are not satisfied with the availability of tee times, we would have entered *No* in cell D4. With *Yes* entered in cell D4, the COUNTIF function in cell D5 counts the number of Yes responses in the sample. The sample proportion is then computed in cell D6 by dividing the number of Yes responses in cell D5 by the sample size in cell D3.

Cells D8:D10 are used to compute the appropriate z value. The confidence coefficient (*0.95*) is entered into cell D8 and the level of significance (α) is computed in cell D9 by entering the formula *=1-D8*. The z value corresponding to an upper-tail area of $\alpha/2$ is computed by entering the formula *=NORM.S.INV(1-D9/2)* into cell D10. The value worksheet shows that $z_{0.025} = 1.96$.

Cells D12:D13 provide the estimate of the standard error and the margin of error. In cell D12, we entered the formula *=SQRT(D6*(1-D6)/D3)* to compute the standard error using the sample proportion and the sample size as inputs. The formula *=D10*D12* is entered into cell D13 to compute the margin of error corresponding to equation (7.9).

Cells D15:D17 provide the point estimate and the lower and upper limits for a confidence interval. The point estimate in cell D15 is the sample proportion. The lower and upper limits in cells D16 and D17 are obtained by subtracting and adding the margin of error to the point estimate. We note that the 95% confidence interval for the proportion of women golfers who are satisfied with the availability of tee times is 0.4076 to 0.4724.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.15 95% Confidence Interval for Survey of Women Golfers

	A	B	C	D	E	F
1	Response		Interval Estimate of a Population Proportion			
2	Yes					
3	No		Sample Size	=COUNTA(A2:A901)		
4	Yes		Response of Interest	Yes		
5	Yes		Count for Response	=COUNTIF(A2:A901,D4)		
6	No		Sample Proportion	=D5/D3		
7	No					
8	No		Confidence Coefficient	0.95		
9	Yes		Level of Significance (alpha)	=1-D8		
10	Yes		z Value	=NORM.S.INV(1-D9/2)		
11	Yes					
12	No		Standard Error	=SQRT(D6*(1-D6)/D3)		
13	No		Margin of Error	=D10*D12		
14	Yes					
15	No		Point Estimate	=D6		
16	No		Lower Limit	=D15-D13		
17	Yes		Upper Limit	=D15+D13		
18	No					
900	Yes					
901	Yes					
902						

Enter Yes as the Response of Interest

Enter 0.95 as the Confidence Coefficient

	A	B	C	D	E
1	Response		Interval Estimate of a Population Proportion		
2	Yes				
3	No		Sample Size	900	
4	Yes		Response of Interest	Yes	
5	Yes		Count for Response	396	
6	No		Sample Proportion	0.44	
7	No				
8	No		Confidence Coefficient	0.95	
9	Yes		Level of Significance (alpha)	0.05	
10	Yes		z Value	1.96	
11	Yes				
12	No		Standard Error	0.0165	
13	No		Margin of Error	0.0324	
14	Yes				
15	No		Point Estimate	0.44	
16	No		Lower Limit	0.4076	
17	Yes		Upper Limit	0.4724	
18	No				
900	Yes				
901	Yes				
902					

Note: Rows 19–899 are hidden.

Notes + Comments

1. The reason the number of degrees of freedom associated with the t value in equation (7.7) is $n - 1$ concerns the use of s as an estimate of the population standard deviation σ. The expression for the sample standard deviation is

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}}.$$

Degrees of freedom refer to the number of independent pieces of information that go into the computation of $\Sigma(x_i - \bar{x})^2$. The n pieces of information involved in computing $\Sigma(x_i - \bar{x})^2$ are as follows: $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$. Note that $\Sigma(x_i - \bar{x}) = 0$ for any data set. Thus, only $n - 1$ of the $x_i - \bar{x}$ values are independent; that is, if we know $n - 1$ of the values, the remaining value can be determined exactly by using the condition that the sum of the $x_i - \bar{x}$ values must be 0. Thus, $n - 1$ is the number of degrees of freedom associated with $\Sigma(x_i - \bar{x})^2$ and hence the number of degrees of freedom for the t distribution in equation (7.7).

2. In most applications, a sample size of $n \geq 30$ is adequate when using equation (7.7) to develop an interval estimate of a population mean. However, if the population distribution is highly skewed or contains outliers, most statisticians would recommend increasing the sample size to 50 or more. If the population is not normally distributed but is roughly symmetric, sample sizes as small as 15 can be expected to provide good approximate confidence intervals. With smaller sample sizes, equation (7.7) should be used if the analyst believes, or is willing to assume, that the population distribution is at least approximately normal.
3. What happens to confidence interval estimates of $\bar{x}$ when the population is skewed? Consider a population that is skewed to the right, with large data values stretching the distribution to the right. When such skewness exists, the sample mean x and the sample standard deviation s are positively correlated. Larger values of s tend to be

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

associated with larger values of $\bar{x}$. Thus, when $\bar{x}$ is larger than the population mean, s tends to be larger than σ. This skewness causes the margin of error, $t_{\alpha/2}(s/\sqrt{n})$, to be larger than it would be with σ known. The confidence interval with the larger margin of error tends to include the population mean more often than it would if the true value of σ were used. But when $\bar{x}$ is smaller than the population mean, the correlation between $\bar{x}$ and s causes the margin of error to be small. In this case, the confidence interval with the smaller margin of error tends to miss the population mean more than it would if we knew σ and used it. For this reason, we recommend using larger sample sizes with highly skewed population distributions.

4. We can find the sample size necessary to provide the desired margin of error at the chosen confidence level. Let E = the desired margin of error. Then
   - the sample size for an interval estimate of a population mean is $n = \dfrac{(z_{\alpha/2})^2\sigma^2}{E^2}$, where E is the margin of error that the user is willing to accept, and the value of $z_{\alpha/2}$ follows directly from the confidence level to be used in developing the interval estimate.
   - the sample size for an interval estimate of a population proportion is $n = \dfrac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2}$, where the planning value p^* can be chosen by use of (i) the sample proportion from a previous sample of the same or similar units, (ii) a pilot study to select a preliminary sample, (iii) judgment or a "best guess" for the value of p^*, or (iv) if none of the preceding alternatives apply, use of the planning value of $p^* = 0.50$.
5. The desired margin of error for estimating a population proportion is almost always 0.10 or less. In national public opinion polls conducted by organizations such as Gallup and Harris, a 0.03 or 0.04 margin of error is common. With such margins of error, the sample found with $n = \dfrac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2}$ will almost always provide a size that is sufficient to satisfy the requirements of $np \geq 5$ and $n(1-p) \geq 5$ for using a normal distribution as an approximation for the sampling distribution of $\bar{p}$.

7.5 Hypothesis Tests

Throughout this chapter, we have shown how a sample could be used to develop point and interval estimates of population parameters such as the mean μ and the proportion p. In this section, we continue the discussion of statistical inference by showing how hypothesis testing can be used to determine whether a statement about the value of a population parameter should or should not be rejected.

In **hypothesis testing**, we begin by making a tentative conjecture about a population parameter. This tentative conjecture is called the **null hypothesis** and is denoted by H_0. We then define another hypothesis, called the **alternative hypothesis**, which is the opposite of what is stated in the null hypothesis. The alternative hypothesis is denoted by H_a. The hypothesis testing procedure uses data from a sample to test the validity of the two competing statements about a population that are indicated by H_0 and H_a.

This section shows how hypothesis tests can be conducted about a population mean and a population proportion. We begin by providing examples that illustrate approaches to developing null and alternative hypotheses.

Developing Null and Alternative Hypotheses

Learning to formulate hypotheses correctly will take some practice. Expect some initial confusion about the proper choice of the null and alternative hypotheses. The examples in this section are intended to provide guidelines.

It is not always obvious how the null and alternative hypotheses should be formulated. Care must be taken to structure the hypotheses appropriately so that the hypothesis testing conclusion provides the information the researcher or decision maker wants. The context of the situation is very important in determining how the hypotheses should be stated. All hypothesis testing applications involve collecting a random sample and using the sample results to provide evidence for drawing a conclusion. Good questions to consider when formulating the null and alternative hypotheses are, What is the purpose of collecting the sample? What conclusions are we hoping to make?

In the introduction to this section, we stated that the null hypothesis H_0 is a tentative conjecture about a population parameter such as a population mean or a population proportion. The alternative hypothesis H_a is a statement that is the opposite of what is stated in the null hypothesis. In some situations it is easier to identify the alternative hypothesis first and then develop the null hypothesis. In other situations, it is easier to identify the null hypothesis first

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

and then develop the alternative hypothesis. We will illustrate these situations in the following examples.

The Alternative Hypothesis as a Research Hypothesis Many applications of hypothesis testing involve an attempt to gather evidence in support of a research hypothesis. In these situations, it is often best to begin with the alternative hypothesis and make it the conclusion that the researcher hopes to support. Consider a particular automobile that currently attains a fuel efficiency of 24 miles per gallon for city driving. A product research group has developed a new fuel injection system designed to increase the miles-per-gallon rating. The group will run controlled tests with the new fuel injection system looking for statistical support for the conclusion that the new fuel injection system provides more miles per gallon than the current system.

Several new fuel injection units will be manufactured, installed in test automobiles, and subjected to research-controlled driving conditions. The sample mean miles per gallon for these automobiles will be computed and used in a hypothesis test to determine whether it can be concluded that the new system provides more than 24 miles per gallon. In terms of the population mean miles per gallon μ, the research hypothesis $\mu > 24$ becomes the alternative hypothesis. Since the current system provides an average or mean of 24 miles per gallon, we will make the tentative conjecture that the new system is no better than the current system and choose $\mu \leq 24$ as the null hypothesis. The null and alternative hypotheses are as follows:

$$H_0: \mu \leq 24$$
$$H_a: \mu > 24$$

The conclusion that the research hypothesis is true is made if the sample data provide sufficient evidence to show that the null hypothesis can be rejected.

If the sample results lead to the conclusion to reject H_0, the inference can be made that $H_a: \mu > 24$ is true. The researchers have the statistical support to state that the new fuel injection system increases the mean number of miles per gallon. The production of automobiles with the new fuel injection system should be considered. However, if the sample results lead to the conclusion that H_0 cannot be rejected, the researchers cannot conclude that the new fuel injection system is better than the current system. Production of automobiles with the new fuel injection system on the basis of better gas mileage cannot be justified. Perhaps more research and further testing can be conducted.

Successful companies stay competitive by developing new products, new methods, and new services that are better than what is currently available. Before adopting something new, it is desirable to conduct research to determine whether there is statistical support for the conclusion that the new approach is indeed better. In such cases, the research hypothesis is stated as the alternative hypothesis. For example, a new teaching method is developed that is believed to be better than the current method. The alternative hypothesis is that the new method is better; the null hypothesis is that the new method is no better than the old method. A new sales force bonus plan is developed in an attempt to increase sales. The alternative hypothesis is that the new bonus plan increases sales; the null hypothesis is that the new bonus plan does not increase sales. A new drug is developed with the goal of lowering blood pressure more than an existing drug. The alternative hypothesis is that the new drug lowers blood pressure more than the existing drug; the null hypothesis is that the new drug does not provide lower blood pressure than the existing drug. In each case, rejection of the null hypothesis H_0 provides statistical support for the research hypothesis. We will see many examples of hypothesis tests in research situations such as these throughout this chapter and in the remainder of the text.

The Null Hypothesis as a Conjecture to Be Challenged Of course, not all hypothesis tests involve research hypotheses. In the following discussion we consider applications of hypothesis testing where we begin with a belief or a conjecture that a statement about the value of a population parameter is true. We will then use a hypothesis test to challenge the conjecture and determine whether there is statistical evidence to conclude that the conjecture is incorrect. In these situations, it is helpful to develop the null hypothesis first. The null hypothesis H_0 expresses the belief or conjecture about the value of the population parameter. The alternative hypothesis H_a is that the belief or conjecture is incorrect.

As an example, consider the situation of a manufacturer of soft drink products. The label on a soft drink bottle states that it contains 67.6 fluid ounces. We consider the label correct provided the population mean filling weight for the bottles is *at least* 67.6 fluid

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

ounces. With no reason to believe otherwise, we would give the manufacturer the benefit of the doubt and assume that the statement provided on the label is correct. Thus, in a hypothesis test about the population mean fluid weight per bottle, we would begin with the conjecture that the label is correct and state the null hypothesis as $\mu \geq 67.6$. The challenge to this conjecture would imply that the label is incorrect and the bottles are being underfilled. This challenge would be stated as the alternative hypothesis $\mu < 67.6$. Thus, the null and alternative hypotheses are as follows:

$$H_0: \mu \geq 67.6$$
$$H_a: \mu < 67.6$$

A manufacturer's product information is usually assumed to be true and stated as the null hypothesis. The conclusion that the information is incorrect can be made if the null hypothesis is rejected.

A government agency with the responsibility for validating manufacturing labels could select a sample of soft drink bottles, compute the sample mean filling weight, and use the sample results to test the preceding hypotheses. If the sample results lead to the conclusion to reject H_0, the inference that $H_a: \mu < 67.6$ is true can be made. With this statistical support, the agency is justified in concluding that the label is incorrect and that the bottles are being underfilled. Appropriate action to force the manufacturer to comply with labeling standards would be considered. However, if the sample results indicate H_0 cannot be rejected, the conjecture that the manufacturer's labeling is correct cannot be rejected. With this conclusion, no action would be taken.

Let us now consider a variation of the soft drink bottle-filling example by viewing the same situation from the manufacturer's point of view. The bottle-filling operation has been designed to fill soft drink bottles with 67.6 fluid ounces as stated on the label. The company does not want to underfill the containers because that could result in complaints from customers or, perhaps, a government agency. However, the company does not want to overfill containers either because putting more soft drink than necessary into the containers would be an unnecessary cost. The company's goal would be to adjust the bottle-filling operation so that the population mean filling weight per bottle is 67.6 fluid ounces as specified on the label.

Although this is the company's goal, from time to time any production process can get out of adjustment. If this occurs in our example, underfilling or overfilling of the soft drink bottles will occur. In either case, the company would like to know about it in order to correct the situation by readjusting the bottle-filling operation to result in the designated 67.6 fluid ounces. In this hypothesis testing application, we would begin with the conjecture that the production process is operating correctly and state the null hypothesis as $\mu = 67.6$ fluid ounces. The alternative hypothesis that challenges this conjecture is that $\mu \neq 67.6$, which indicates that either overfilling or underfilling is occurring. The null and alternative hypotheses for the manufacturer's hypothesis test are as follows:

$$H_0: \mu = 67.6$$
$$H_a: \mu \neq 67.6$$

Suppose that the soft drink manufacturer uses a quality-control procedure to periodically select a sample of bottles from the filling operation and computes the sample mean filling weight per bottle. If the sample results lead to the conclusion to reject H_0, the inference is made that $H_a: \mu \neq 67.6$ is true. We conclude that the bottles are not being filled properly and the production process should be adjusted to restore the population mean to 67.6 fluid ounces per bottle. However, if the sample results indicate H_0 cannot be rejected, the conjecture that the manufacturer's bottle-filling operation is functioning properly cannot be rejected. In this case, no further action would be taken and the production operation would continue to run.

The two preceding forms of the soft drink manufacturing hypothesis test show that the null and alternative hypotheses may vary depending on the point of view of the researcher or decision maker. To formulate hypotheses correctly, it is important to understand the context of the situation and to structure the hypotheses to provide the information the researcher or decision maker wants.

Summary of Forms for Null and Alternative Hypotheses The hypothesis tests in this chapter involve two population parameters: the population mean and the population proportion. Depending on the situation, hypothesis tests about a population parameter may

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

take one of three forms: Two use inequalities in the null hypothesis; the third uses an equality in the null hypothesis. For hypothesis tests involving a population mean, we let μ_0 denote the hypothesized value of the population mean and we must choose one of the following three forms for the hypothesis test:

The three possible forms of hypotheses H_0 and H_a are shown here. Note that the equality always appears in the null hypothesis H_0.

$$H_0: \mu \geq \mu_0 \quad H_0: \mu \leq \mu_0 \quad H_0: \mu = \mu_0$$
$$H_a: \mu < \mu_0 \quad H_a: \mu > \mu_0 \quad H_a: \mu \neq \mu_0$$

For reasons that will be clear later, the first two forms are called one-tailed tests. The third form is called a two-tailed test.

In many situations, the choice of H_0 and H_a is not obvious and judgment is necessary to select the proper form. However, as the preceding forms show, the equality part of the expression (either $\geq$, $\leq$, or $=$) *always* appears in the null hypothesis. In selecting the proper form of H_0 and H_a, keep in mind that the alternative hypothesis is often what the test is attempting to establish. Hence, asking whether the user is looking for evidence to support $\mu < \mu_0$, $\mu > \mu_0$, or $\mu \neq \mu_0$ will help determine H_a.

Type I and Type II Errors

The null and alternative hypotheses are competing statements about the population. Either the null hypothesis H_0 is true or the alternative hypothesis H_a is true, but not both. Ideally the hypothesis testing procedure should lead to the acceptance of H_0 when H_0 is true and the rejection of H_0 when H_a is true. Unfortunately, the correct conclusions are not always possible. Because hypothesis tests are based on sample information, we must allow for the possibility of errors. Table 7.6 illustrates the two kinds of errors that can be made in hypothesis testing.

The first row of Table 7.6 shows what can happen if the conclusion is to accept H_0. If H_0 is true, this conclusion is correct. However, if H_a is true, we made a **Type II error**; that is, we accepted H_0 when it is false. The second row of Table 7.6 shows what can happen if the conclusion is to reject H_0. If H_0 is true, we made a **Type I error**; that is, we rejected H_0 when it is true. However, if H_a is true, rejecting H_0 is correct.

Recall the hypothesis testing illustration in which an automobile product research group developed a new fuel injection system designed to increase the miles-per-gallon rating of a particular automobile. With the current model obtaining an average of 24 miles per gallon, the hypothesis test was formulated as follows:

$$H_0: \mu \leq 24$$
$$H_a: \mu > 24$$

The alternative hypothesis, $H_a: \mu > 24$, indicates that the researchers are looking for sample evidence to support the conclusion that the population mean miles per gallon with the new fuel injection system is greater than 24.

In this application, the Type I error of rejecting H_0 when it is true corresponds to the researchers claiming that the new system improves the miles-per-gallon rating ($\mu > 24$) when in fact the new system is no better than the current system. In contrast, the Type II

Table 7.6 Errors and Correct Conclusions in Hypothesis Testing

		Population Condition	
		H_0 True	H_a True
Conclusion	Do Not Reject H_0	Correct conclusion	Type II error
	Reject H_0	Type I error	Correct conclusion

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

error of accepting H_0 when it is false corresponds to the researchers concluding that the new system is no better than the current system ($\mu \leq 24$) when in fact the new system improves miles-per-gallon performance.

For the miles-per-gallon rating hypothesis test, the null hypothesis is H_0: $\mu \leq 24$. Suppose the null hypothesis is true as an equality; that is, $\mu = 24$. The probability of making a Type I error when the null hypothesis is true as an equality is called the level of significance. Thus, for the miles-per-gallon rating hypothesis test, the level of significance is the probability of rejecting H_0: $\mu \leq 24$ when $\mu = 24$. Because of the importance of this concept, we now restate the definition of level of significance.

The Greek symbol α (alpha) is used to denote the level of significance, and common choices for α are 0.05 and 0.01.

Level of Significance

The level of significance is the probability of making a Type I error when the null hypothesis is true as an equality.

In practice, the person responsible for the hypothesis test specifies the level of significance. By selecting α, that person is controlling the probability of making a Type I error. If the cost of making a Type I error is high, small values of α are preferred. If the cost of making a Type I error is not too high, larger values of α are typically used. Applications of hypothesis testing that only control the Type I error are called *significance tests*. Many applications of hypothesis testing are of this type.

Although most applications of hypothesis testing control the probability of making a Type I error, they do not always control the probability of making a Type II error. Hence, if we decide to accept H_0, we cannot determine how confident we can be with that decision. Because of the uncertainty associated with making a Type II error when conducting significance tests, statisticians usually recommend that we use the statement "do not reject H_0" instead of "accept H_0." Using the statement "do not reject H_0" carries the recommendation to withhold both judgment and action. In effect, by not directly accepting H_0, the statistician avoids the risk of making a Type II error. Whenever the probability of making a Type II error has not been determined and controlled, we will not make the statement "accept H_0." In such cases, only two conclusions are possible: *do not reject* H_0 or *reject* H_0.

If the sample data are consistent with the null hypothesis H_0, we will follow the practice of concluding "do not reject H_0." This conclusion is preferred over "accept H_0," because the conclusion to accept H_0 puts us at risk of making a Type II error.

Although controlling for a Type II error in hypothesis testing is not common, it can be done. Specialized texts describe procedures for determining and controlling the probability of making a Type II error.[1] If proper controls have been established for this error, action based on the "accept H_0" conclusion can be appropriate.

Hypothesis Test of the Population Mean

In this section, we describe how to conduct hypothesis tests about a population mean for the practical situation in which the sample must be used to develop estimates of both μ and σ. Thus, to conduct a hypothesis test about a population mean, the sample mean $\bar{x}$ is used as an estimate of μ and the sample standard deviation s is used as an estimate of σ.

One-Tailed Test **One-tailed tests** about a population mean take one of the following two forms:

Lower-Tail Test	Upper-Tail Test
H_0: $\mu \geq \mu_0$	H_0: $\mu \leq \mu_0$
H_a: $\mu < \mu_0$	H_a: $\mu > \mu_0$

Let us consider an example involving a lower-tail test.

[1]See, for example, D. R. Anderson, D. J. Sweeney, T. A. Williams, J. D. Camm, J. J. Cochran, M. J. Fry, and J. W. Ohlmann *Statistics for Business and Economics*, 14th ed. (Mason, OH: Cengage Learning, 2020).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The Federal Trade Commission (FTC) periodically conducts statistical studies designed to test the claims that manufacturers make about their products. For example, the label on a large can of Hilltop Coffee states that the can contains 3 pounds of coffee. The FTC knows that Hilltop's production process cannot place exactly 3 pounds of coffee in each can, even if the mean filling weight for the population of all cans filled is 3 pounds per can. However, as long as the population mean filling weight is at least 3 pounds per can, the rights of consumers will be protected. Thus, the FTC interprets the label information on a large can of coffee as a claim by Hilltop that the population mean filling weight is at least 3 pounds per can. We will show how the FTC can check Hilltop's claim by conducting a lower-tail hypothesis test.

The first step is to develop the null and alternative hypotheses for the test. If the population mean filling weight is at least 3 pounds per can, Hilltop's claim is correct. This establishes the null hypothesis for the test. However, if the population mean weight is less than 3 pounds per can, Hilltop's claim is incorrect. This establishes the alternative hypothesis. With μ denoting the population mean filling weight, the null and alternative hypotheses are as follows:

$$H_0: \mu \geq 3$$
$$H_a: \mu < 3$$

Note that the hypothesized value of the population mean is $\mu_0 = 3$.

If the sample data indicate that H_0 cannot be rejected, the statistical evidence does not support the conclusion that a label violation has occurred. Hence, no action should be taken against Hilltop. However, if the sample data indicate that H_0 can be rejected, we will conclude that the alternative hypothesis, $H_a: \mu < 3$, is true. In this case a conclusion of underfilling and a charge of a label violation against Hilltop would be justified.

DATA*file*
coffee

Suppose a sample of 36 cans of coffee is selected and the sample mean $\bar{x}$ is computed as an estimate of the population mean μ. If the value of the sample mean $\bar{x}$ is less than 3 pounds, the sample results will cast doubt on the null hypothesis. What we want to know is how much less than 3 pounds must $\bar{x}$ be before we would be willing to declare the difference significant and risk making a Type I error by falsely accusing Hilltop of a label violation. A key factor in addressing this issue is the value the decision maker selects for the level of significance.

As noted in the preceding section, the level of significance, denoted by α, is the probability of making a Type I error by rejecting H_0 when the null hypothesis is true as an equality. The decision maker must specify the level of significance. If the cost of making a Type I error is high, a small value should be chosen for the level of significance. If the cost is not high, a larger value is more appropriate. In the Hilltop Coffee study, the director of the FTC's testing program made the following statement: "If the company is meeting its weight specifications at $\mu = 3$, I do not want to take action against them. But I am willing to risk a 1% chance of making such an error." From the director's statement, we set the level of significance for the hypothesis test at $\alpha = 0.01$. Thus, we must design the hypothesis test so that the probability of making a Type I error when $\mu = 3$ is 0.01.

For the Hilltop Coffee study, by developing the null and alternative hypotheses and specifying the level of significance for the test, we carry out the first two steps required in conducting every hypothesis test. We are now ready to perform the third step of hypothesis testing: collect the sample data and compute the value of what is called a test statistic.

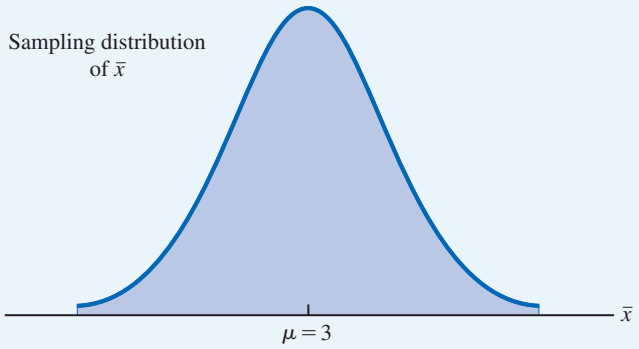
The standard error of $\bar{x}$ is the standard deviation of the sampling distribution of $\bar{x}$.

Test Statistic From the study of sampling distributions in Section 7.3 we know that as the sample size increases, the sampling distribution of $\bar{x}$ will become normally distributed. Figure 7.16 shows the sampling distribution of $\bar{x}$ when the null hypothesis is true as an equality, that is, when $\mu = \mu_0 = 3$.[2] Note that $\sigma_{\bar{x}}$, the standard error of $\bar{x}$, is estimated by

[2]In constructing sampling distributions for hypothesis tests, it is assumed that H_0 is satisfied as an equality.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.16 Sampling Distribution of $\bar{x}$ for the Hilltop Coffee Study When the Null Hypothesis Is True as an Equality ($\mu = 3$)



$s_{\bar{x}} = s/\sqrt{n} = 0.17\sqrt{36} = 0.028$. Recall that in Section 7.4, we showed that an interval estimate of a population mean is based on a probability distribution known as the t distribution. The t distribution is similar to the standard normal distribution, but accounts for the additional variability introduced when using a sample to estimate both the population mean and population standard deviation. Hypothesis tests about a population mean are also based on the t distribution. Specifically, if $\bar{x}$ is normally distributed, the sampling distribution of

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\bar{x} - 3}{0.028}$$

is a t distribution with $n - 1$ degrees of freedom. The value of t represents how much the sample mean is above or below the hypothesized value of the population mean as measured in units of the standard error of the sample mean. A value of $t = -1$ means that the value of $\bar{x}$ is 1 standard error below the hypothesized value of the mean, a value of $t = -2$ means that the value of $\bar{x}$ is 2 standard errors below the hypothesized value of the mean, and so on. For this lower-tail hypothesis test, we can use Excel to find the lower-tail probability corresponding to any t value (as we show later in this section). For example, Figure 7.17 illustrates that the lower tail area at $t = -3.00$ is 0.0025. Hence, the probability of obtaining a value of t that is three or more standard errors below the mean is 0.0025. As a result, if the null hypothesis is true (i.e., if the population mean is 3), the probability of obtaining a value of $\bar{x}$ that is 3 or more standard errors below the hypothesized population mean $\mu_0 = 3$ is also 0.0025. Because such a result is unlikely if the null hypothesis is true, this leads us to doubt our null hypothesis.

Although the t *distribution is based on an conjecture that the population from which we are sampling is normally distributed, research shows that when the sample size is large enough, this conjecture can be relaxed considerably.*

We use the t-distributed random variable t as a **test statistic** to determine whether $\bar{x}$ deviates from the hypothesized value of μ enough to justify rejecting the null hypothesis. With $s_{\bar{x}} = s/\sqrt{n}$, the test statistic is as follows:

Test Statistic for Hypothesis Tests About a Population Mean

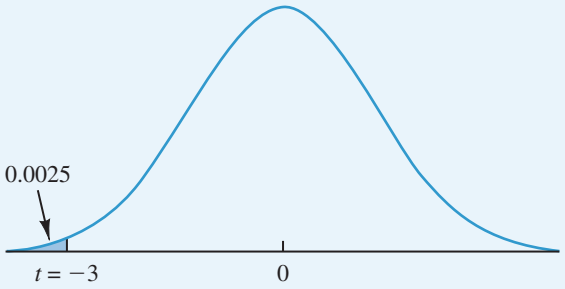
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{7.11}$$

A small p *value indicates that the value of the test statistic is unusual given the conjecture that* H_0 *is true.*

The key question for a lower-tail test is, How small must the test statistic t be before we choose to reject the null hypothesis? We will draw our conclusion by using the value of the test statistic t to compute a probability called a ***p* value**.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.17 Lower-Tail Probability for $t = -3$ from a t Distribution with 35 Degrees of Freedom



p Value

A p value is the probability, assuming that H_0 is true, of obtaining a random sample of size n that results in a test statistic at least as extreme as the one observed in the current sample.

The p value measures the strength of the evidence provided by the sample against the null hypothesis. Smaller p values indicate more evidence against H_0 as they suggest that it is increasingly more unlikely that the sample could occur if the H_0 is true.

Let us see how the p value is computed and used. The value of the test statistic is used to compute the p value. The method used depends on whether the test is a lower-tail, an upper-tail, or a two-tailed test. For a lower-tail test, the p value is the probability of obtaining a value for the test statistic as small as or smaller than that provided by the sample. Thus, to compute the p value for the lower-tail test, we must use the t distribution to find the probability that t is less than or equal to the value of the test statistic. After computing the p value, we must then decide whether it is small enough to reject the null hypothesis; as we will show, this decision involves comparing the p value to the level of significance.

Using Excel Excel can be used to conduct one-tailed and two-tailed hypothesis tests about a population mean. The sample data and the test statistic (t) are used to compute three p values: p value (lower tail), p value (upper tail), and p value (two tail). The user can then choose α and draw a conclusion using whichever p value is appropriate for the type of hypothesis test being conducted.

Let's start by showing how to use Excel's T.DIST function to compute a lower-tail p value. The T.DIST function has three inputs; its general form is as follows:

=T.DIST(test statistic, degrees of freedom, cumulative).

For the first input, we enter the value of the test statistic; for the second input we enter the degrees of freedom for the associated t distribution; for the third input, we enter *TRUE* to compute the cumulative probability corresponding to a lower-tail p value.

Once the lower-tail p value has been computed, it is easy to compute the upper-tail and the two-tailed p values. The upper-tail p value is 1 minus the lower-tail p value, and the two-tailed p value is two times the smaller of the lower- and upper-tail p values.

MODEL file
coffeetest

Let us now compute the p value for the Hilltop Coffee lower-tail test. Refer to Figure 7.18 as we describe the tasks involved. The formula sheet is in the background and the value worksheet is in the foreground.

The descriptive statistics needed are provided in cells D4:D6. Excel's COUNT, AVERAGE, and STDEV.S functions compute the sample size, the sample mean, and the sample standard deviation, respectively. The hypothesized value of the population mean

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.18 Hypothesis Test About a Population Mean

	A	B	C	D
1	Weight		Hypothesis Test about a Population Mean	
2	3.15			
3	2.76			
4	3.18		Sample Size	=COUNT(A2:A37)
5	2.77		Sample Mean	=AVERAGE(A2:A37)
6	2.86		Sample Standard Deviation	=STDEV.S(A2:A37)
7	2.66			
8	2.86		Hypothesized Value	3
9	2.54			
10	3.02		Standard Error	=D6/SQRT(D4)
11	3.13		Test Statistic *t*	=(D5-D8)/D10
12	2.94		Degrees of Freedom	=D4-1
13	2.74			
14	2.84		*p*-value (Lower Tail)	=T.DIST(D11,D12,TRUE)
15	2.6		*p*-value (Upper Tail)	=1-D14
16	2.94		*p*-value (Two Tail)	=2*MIN(D14,D15)

	A	B	C	D
1	Weight		Hypothesis Test about a Population Mean	
2	3.15			
3	2.76			
4	3.18		Sample Size	36
5	2.77		Sample Mean	2.92
6	2.86		Sample Standard Deviation	0.170
7	2.66			
8	2.86		Hypothesized Value	3
9	2.54			
10	3.02		Standard Error	0.028
11	3.13		Test Statistic *t*	-2.824
12	2.94		Degrees of Freedom	35
13	2.74			
14	2.84		*p*-value (Lower Tail)	0.0039
15	2.60		*p*-value (Upper Tail)	0.9961
16	2.94		*p*-value (Two Tail)	0.0078

Note: Rows 17–37 are hidden.

(*3*) is entered into cell D8. Using the sample standard deviation as an estimate of the population standard deviation, an estimate of the standard error is obtained in cell D10 by dividing the sample standard deviation in cell D6 by the square root of the sample size in cell D4. The formula *=(D5-D8)/D10* entered into cell D11 computes the value of the test statistic t corresponding to the calculation:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.92 - 3}{0.17/\sqrt{36}} = -2.824$$

The degrees of freedom are computed in cell D12 as the sample size in cell D4 minus 1.

To compute the p value for a lower-tail test, we enter the following formula into cell D14.

=T.DIST(D11,D12,TRUE)

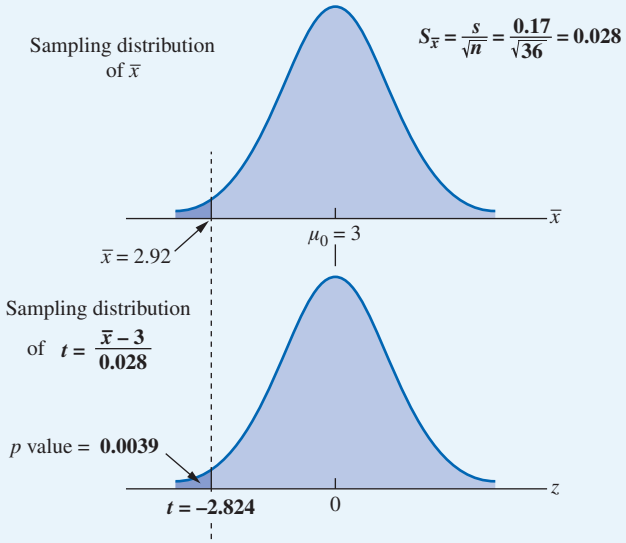
The p value for an upper-tail test is then computed in cell D15 as 1 minus the p value for the lower-tail test. Finally, the p value for a two-tailed test is computed in cell D16 as two times the minimum of the two one-tailed p values. The value worksheet shows that the three p values are p value (lower tail) = 0.0039, p value (upper tail) = 0.9961, and p value (two tail) = 0.0078.

The development of the worksheet is now complete. Is $\bar{x} = 2.92$ small enough to lead us to reject H_0? Because this is a lower-tail test, the p value is the area under the t-distribution curve for values of $t \leq -2.824$ (the value of the test statistic). Figure 7.19 depicts the p value for the Hilltop Coffee lower-tail test. This p value indicates a small probability of obtaining a sample mean of $\bar{x} = 2.92$ (and a test statistic of -2.824) or smaller when sampling from a population with $\mu = 3$. This p value does not provide much support for the null hypothesis, but is it small enough to cause us to reject H_0? The answer depends on the level of significance (α) the decision maker has selected for the test.

Note that the p value can be considered a measure of the strength of the evidence against the null hypothesis that is contained in the sample data. The greater the inconsistency between the sample data and the null hypothesis, the smaller the p value will be; thus, a smaller p value indicates that it is less plausible that the sample could have been collected from a population for which the null hypothesis is true. That is, a smaller p value indicates that the sample provides stronger evidence against the null hypothesis.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.19 *p* Value for the Hilltop Coffee Study When $\bar{x} = 2.92$ and $s = 0.17$



As noted previously, the director of the FTC's testing program selected a value of 0.01 for the level of significance. The selection of $\alpha = 0.01$ means that the director is willing to tolerate a probability of 0.01 of rejecting the null hypothesis when it is true as an equality ($\mu_0 = 3$). The sample of 36 coffee cans in the Hilltop Coffee study resulted in a p value of 0.0039, which means that the probability of obtaining a value of $\bar{x} = 2.92$ or less when the null hypothesis is true is 0.0039. Because 0.0039 is less than or equal to $\alpha = 0.01$, we reject H_0. Therefore, we find sufficient statistical evidence to reject the null hypothesis at the 0.01 level of significance.

The level of significance α indicates the strength of evidence that is needed in the sample data before we will reject the null hypothesis. If the p value is smaller than the selected level of significance α, the evidence against the null hypothesis that is contained in the sample data is sufficiently strong for us to reject the null hypothesis; that is, we believe that it is implausible that the sample data were collected from a population for which H_0: $\mu \geq 3$ is true. Conversely, if the p value is larger than the selected level of significance α, the evidence against the null hypothesis that is contained in the sample data is not sufficiently strong for us to reject the null hypothesis; that is, we believe that it is plausible that the sample data were collected from a population for which the null hypothesis is true.

We can now state the general rule for determining whether the null hypothesis can be rejected when using the p value approach. For a level of significance α, the rejection rule using the p value approach is as follows.

Rejection Rule

Reject H_0 if p value $\leq \alpha$

In the Hilltop Coffee test, the p value of 0.0039 resulted in the rejection of the null hypothesis. Although the basis for making the rejection decision involves a comparison of the p value to the level of significance specified by the FTC director, the observed p value of 0.0039 means that we would reject H_0 for any value of $\alpha \geq 0.0039$. For this reason, the p value is also called the *observed level of significance*.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Different decision makers may express different opinions concerning the cost of making a Type I error and may choose a different level of significance. By providing the p value as part of the hypothesis testing results, another decision maker can compare the reported p value to their own level of significance and possibly make a different decision with respect to rejecting H_0.

Lower-Tail Test	Upper-Tail Test
$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$
$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$

At the beginning of this section, we said that one-tailed tests about a population mean take one of the following two forms:
We used the Hilltop Coffee study to illustrate how to conduct a lower-tail test. We can use the same general approach to conduct an upper-tail test. The test statistic t is still computed using equation (7.11). But, for an upper-tail test, the p value is the probability of obtaining a value for the test statistic as large as or larger than that provided by the sample. Thus, to compute the p value for the upper-tail test, we must use the t distribution to compute the probability that t is greater than or equal to the value of the test statistic. Then, according to the rejection rule, we will reject the null hypothesis if the p value is less than or equal to the level of significance α.

Computation of *p* Values for One-Tailed Tests

1. Compute the value of the test statistic using equation (7.11).
2. **Lower-tail test:** Using the t distribution, compute the probability that t is less than or equal to the value of the test statistic (area in the lower tail).
3. **Upper-tail test:** Using the t distribution, compute the probability that t is greater than or equal to the value of the test statistic (area in the upper tail).

Let us summarize the steps involved in computing p values for one-tailed hypothesis tests.

Two-Tailed Test In hypothesis testing, the general form for a **two-tailed test** about a population mean is as follows:

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

In this subsection we show how to conduct a two-tailed test about a population mean. As an illustration, we consider the hypothesis testing situation facing Holiday Toys.

DATA*file*
orders

Holiday Toys manufactures and distributes its products through more than 1,000 retail outlets. In planning production levels for the coming winter season, Holiday must decide how many units of each product to produce before the actual demand at the retail level is known. For this year's most important new toy, Holiday's marketing director is expecting demand to average 40 units per retail outlet. Prior to making the final production decision based on this estimate, Holiday decided to survey a sample of 25 retailers to gather more information about demand for the new product. Each retailer was provided with information about the features of the new toy along with the cost and the suggested selling price. Then each retailer was asked to specify an anticipated order quantity.

With μ denoting the population mean order quantity per retail outlet, the sample data will be used to conduct the following two-tailed hypothesis test:

$$H_0: \mu = 40$$
$$H_a: \mu \neq 40$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If H_0 cannot be rejected, Holiday will continue its production planning based on the marketing director's estimate that the population mean order quantity per retail outlet will be $\mu = 40$ units. However, if H_0 is rejected, Holiday will immediately reevaluate its production plan for the product. A two-tailed hypothesis test is used because Holiday wants to reevaluate the production plan regardless of whether the population mean quantity per retail outlet is less than anticipated or is greater than anticipated. Because it's a new product and therefore, no historical data are available, the population mean μ and the population standard deviation must both be estimated using $\bar{x}$ and s from the sample data.

The sample of 25 retailers provided a mean of $\bar{x} = 37.4$ and a standard deviation of $s = 11.79$ units. Before going ahead with the use of the t distribution, the analyst constructed a histogram of the sample data in order to check on the form of the population distribution. The histogram of the sample data showed no evidence of skewness or any extreme outliers, so the analyst concluded that the use of the t distribution with $n - 1 = 24$ degrees of freedom was appropriate. Using equation (9.2) with $\bar{x} = 37.4$, $\mu_0 = 40$, $s = 11.79$, and $n = 25$, the value of the test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{37.4 - 40}{11.79/\sqrt{25}} = -1.10$$

The sample mean $\bar{x} = 37.4$ is less than 40 and so provides some support for the conclusion that the population mean quantity per retail outlet is less than 40 units, but this could possibly be due to sampling error. We must address whether the difference between this sample mean and our hypothesized mean is sufficient for us to reject H_0 at the 0.05 level of significance. We will again reach our conclusion by calculating a p value.

Recall that the p value is a probability used to determine whether the null hypothesis should be rejected. For a two-tailed test, values of the test statistic in either tail provide evidence against the null hypothesis. For a two-tailed test the p value is the probability of obtaining a value for the test statistic *at least as unlikely* as the value of the test statistic calculated with the sample given that the null hypothesis is true. Let us see how the p value is computed for the two-tailed Holiday Toys hypothesis test.

To compute the p value for this problem, we must find the probability of obtaining a value for the test statistic at least as unlikely as $t = -1.10$ if the population mean is actually 40. Clearly, values of $t \leq -1.10$ are *at least as unlikely*. But because this is a two-tailed test, all values that are more than 1.10 standard deviations from the hypothesized value μ_0 in either direction provide evidence against the null hypothesis that is at least as strong as the evidence against the null hypothesis contained in the sample data. As shown in Figure 7.20, the two-tailed p value in this case is given by $P(t \leq -1.10) + P(t \geq 1.10)$.

To compute the tail probabilities, we apply the Excel template introduced in the Hilltop Coffee example to the Holiday Toys data. Figure 7.21 displays the formula worksheet in the background and the value worksheet in the foreground.

To complete the two-tailed Holiday Toys hypothesis test, we compare the two-tailed p value to the level of significance to see whether the null hypothesis should be rejected. With a level of significance of $\alpha = 0.05$, we do not reject H_0 because the two-tailed p value $= 0.2811 > 0.05$. This result indicates that Holiday should continue its production planning for the coming season based on the expectation that $\mu = 40$.

MODEL file
orderstest

The *orderstest* worksheet in Figure 7.21 can be used as a template for any hypothesis tests about a population mean. To facilitate the use of this worksheet, the formulas in cells D4:D6 reference the entire column A as follows:

Cell D4: =*COUNT(A:A)*
Cell D5: =*AVERAGE(A:A)*
Cell D6: =*STDEV.S(A:A)*

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 7.20 *p* Value for the Holiday Toys Two-Tailed Hypothesis Test

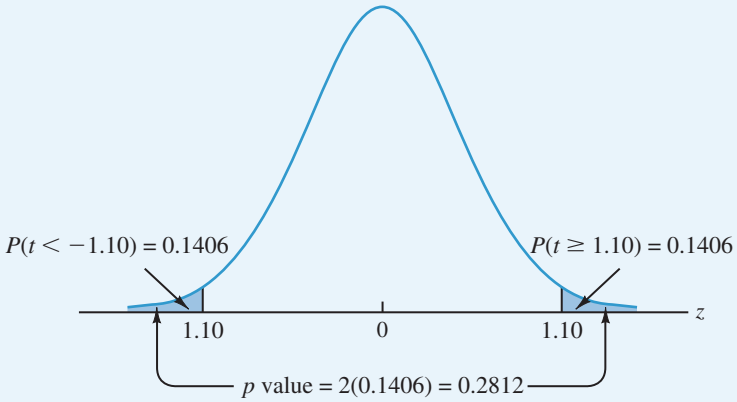


Figure 7.21 Two-Tailed Hypothesis Test for Holiday Toys

	A	B	C	D
1	Units		Hypothesis Test about a Population Mean	
2	26			
3	23			
4	32		Sample Size	=COUNT(A:A)
5	47		Sample Mean	=AVERAGE(A:A)
6	45		Sample Standard Deviation	=STDEV.S(A:A)
7	31			
8	47		Hypothesized Value	40
9	59			
10	21		Standard Error	=D6/SQRT(D4)
11	52		Test Statistic *t*	=(D5-D8)/D10
12	45		Degrees of Freedom	=D4-1
13	53			
14	34		*p*-value (Lower Tail)	=T.DIST(D11,D12,TRUE)
15	45		*p*-value (Upper Tail)	=1-D14
16	39		*p*-value (Two Tail)	=2*MIN(D14,D15)

	A	B	C	D
1	Units		Hypothesis Test about a Population Mean	
2	26			
3	23			
4	32		Sample Size	25
5	47		Sample Mean	37.4
6	45		Sample Standard Deviation	11.79
7	31			
8	47		Hypothesized Value	40
9	59			
10	21		Standard Error	2.358
11	52		Test Statistic *t*	-1.103
12	45		Degrees of Freedom	24
13	53			
14	34		*p*-value (Lower Tail)	0.1406
15	45		*p*-value (Upper Tail)	0.8594
16	39		*p*-value (Two Tail)	0.2811

Note: Rows 17–26 are hidden.

With the A:A method of specifying data ranges, Excel's COUNT function will count the number of numeric values in column A, Excel's AVERAGE function will compute the average of the numeric values in column A, and Excel's STDEV.S function will compute the standard deviation of the numeric values in Column A. Thus, to solve a new problem it is necessary only to enter the new data in column A and enter the hypothesized value of the population mean in cell D8. Then, the standard error, the test statistic, degrees of freedom, and the three *p* values will be updated by the Excel formulas.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Let us summarize the steps involved in computing p values for two-tailed hypothesis tests.

Computation of p Values for Two-Tailed Tests

1. Compute the value of the test statistic using equation (7.11).
2. If the value of the test statistic is in the upper tail, compute the probability that t is greater than or equal to the value of the test statistic (the upper-tail area). If the value of the test statistic is in the lower tail, compute the probability that t is less than or equal to the value of the test statistic (the lower-tail area).
3. Double the probability (or tail area) from step 2 to obtain the p value.

Summary and Practical Advice We presented examples of a lower-tail test and a two-tailed test about a population mean. Based on these examples, we can now summarize the hypothesis testing procedures about a population mean in Table 7.7. Note that μ_0 is the hypothesized value of the population mean.

The hypothesis testing steps followed in the two examples presented in this section are common to every hypothesis test.

Steps of Hypothesis Testing

Step 1. Develop the null and alternative hypotheses.
Step 2. Specify the level of significance.
Step 3. Collect the sample data and compute the value of the test statistic.
Step 4. Use the value of the test statistic to compute the p value.
Step 5. Reject H_0 if the $p \leq \alpha$.
Step 6. Interpret the statistical conclusion in the context of the application.

Practical advice about the sample size for hypothesis tests is similar to the advice we provided about the sample size for interval estimation in Section 7.4. In most applications, a sample size of $n \geq 30$ is adequate when using the hypothesis testing procedure described in this section. In cases in which the sample size is less than 30, the distribution of the population from which we are sampling becomes an important consideration. When the population is normally distributed, the hypothesis tests described in this section provide exact results for any sample size. When the population is not normally distributed, these procedures provide approximations. Nonetheless, we find that sample sizes of 30 or more will provide good results in most cases. If the population is approximately normal, small sample sizes (e.g., $n = 15$) can provide acceptable results. If the population is highly skewed or contains outliers, sample sizes approaching 50 are recommended.

Table 7.7 Summary of Hypothesis Tests About a Population Mean

	Lower-Tail Test	Upper-Tail Test	Two-Tailed Test
Hypotheses	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Test Statistic	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
***p* Value**	=T.DIST(t,n − 1, TRUE)	=1 − T.DIST(t,n − 1, TRUE)	=2*MIN(T.DIST(t,n − 1, TRUE),1 − T.DIST(t,n − 1, TRUE))

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Relationship Between Interval Estimation and Hypothesis Testing In Section 7.4 we showed how to develop a confidence interval estimate of a population mean. The $(1 - \alpha)\%$ confidence interval estimate of a population mean is given by

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

In this chapter we showed that a two-tailed hypothesis test about a population mean takes the following form:

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

where μ_0 is the hypothesized value for the population mean.

Suppose that we follow the procedure described in Section 7.4 for constructing a $100(1 - \alpha)\%$ confidence interval for the population mean. We know that $100(1 - \alpha)\%$ of the confidence intervals generated will contain the population mean and $100\alpha\%$ of the confidence intervals generated will not contain the population mean. Thus, if we reject H_0 whenever the confidence interval does not contain μ_0, we will be rejecting the null hypothesis when it is true ($\mu = \mu_0$) with probability α. Recall that the level of significance is the probability of rejecting the null hypothesis when it is true. So constructing a $100(1 - \alpha)\%$ confidence interval and rejecting H_0 whenever the interval does not contain μ_0 is equivalent to conducting a two-tailed hypothesis test with α as the level of significance. The procedure for using a confidence interval to conduct a two-tailed hypothesis test can now be summarized.

A Confidence Interval Approach to Testing a Hypothesis of the Form

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

1. Select a simple random sample from the population and use the value of the sample mean $\bar{x}$ to develop the confidence interval for the population mean μ.

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

2. If the confidence interval contains the hypothesized value μ_0, do not reject H_0. Otherwise, reject[3] H_0.

For a two-tailed hypothesis test, the null hypothesis can be rejected if the confidence interval does not include μ_0.

Let us illustrate by conducting the Holiday Toys hypothesis test using the confidence interval approach. The Holiday Toys hypothesis test takes the following form:

$$H_0: \mu = 40$$
$$H_a: \mu \neq 40$$

To test these hypotheses with a level of significance of $\alpha = 0.05$, we sampled 25 retailers and found a sample mean of $\bar{x} = 37.4$ units and a sample standard deviation of $s = 11.79$ units. Using these results with $t_{0.025} = \text{T.INV}(1 - (.05/2), 25 - 1) = 2.064$, we find that the 95% confidence interval estimate of the population mean is

$$\bar{x} \pm t_{0.025} \frac{s}{\sqrt{n}}$$
$$37.4 \pm 2.064 \frac{11.79}{\sqrt{25}}$$
$$37.4 \pm 4.4$$

or

$$33.0 \text{ to } 41.8.$$

This finding enables Holiday's marketing director to conclude with 95% confidence that the mean number of units per retail outlet is between 33.0 and 41.8. Because the

[3]To be consistent with the rule for rejecting H_0 when $p \leq \alpha$, we would also reject H_0 using the confidence interval approach if μ_0 happens to be equal to one of the endpoints of the $100(1 - \alpha)\%$ confidence interval.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

hypothesized value for the population mean, $\mu_0 = 40$, is in this interval, the hypothesis testing conclusion is that the null hypothesis, $H_0: \mu = 40$, cannot be rejected.

Note that this discussion and example pertain to two-tailed hypothesis tests about a population mean. However, the same confidence interval and two-tailed hypothesis testing relationship exists for other population parameters. The relationship can also be extended to one-tailed tests about population parameters. Doing so, however, requires the development of one-sided confidence intervals, which are rarely used in practice.

Hypothesis Test of the Population Proportion

In this section we show how to conduct a hypothesis test about a population proportion p. Using p_0 to denote the hypothesized value for the population proportion, the three forms for a hypothesis test about a population proportion are as follows:

$$H_0: p \geq p_0 \quad H_0: p \leq p_0 \quad H_0: p = p_0$$
$$H_a: p < p_0 \quad H_a: p > p_0 \quad H_a: p \neq p_0$$

The first form is called a lower-tail test, the second an upper-tail test, and the third form a two-tailed test.

Hypothesis tests about a population proportion are based on the difference between the sample proportion $\bar{p}$ and the hypothesized population proportion p_0. The methods used to conduct the hypothesis test are similar to those used for hypothesis tests about a population mean. The only difference is that we use the sample proportion and its standard error to compute the test statistic. The p value is then used to determine whether the null hypothesis should be rejected.

Let us consider an example involving a situation faced by Pine Creek golf course. Over the past year, 20% of the players at Pine Creek were women. In an effort to increase the proportion of women players, Pine Creek implemented a special promotion designed to attract women golfers. One month after the promotion was implemented, the course manager requested a statistical study to determine whether the proportion of women players at Pine Creek had increased. Because the objective of the study is to determine whether the proportion of women golfers increased, an upper-tail test with $H_a: p > 0.20$ is appropriate. The null and alternative hypotheses for the Pine Creek hypothesis test are as follows:

$$H_0: p \leq 0.20$$
$$H_a: p > 0.20$$

If H_0 can be rejected, the test results will give statistical support for the conclusion that the proportion of women golfers increased and the promotion was beneficial. The course manager specified that a level of significance of $\alpha = 0.05$ be used in carrying out this hypothesis test.

The next step of the hypothesis testing procedure is to select a sample and compute the value of an appropriate test statistic. To show how this step is done for the Pine Creek upper-tail test, we begin with a general discussion of how to compute the value of the test statistic for any form of a hypothesis test about a population proportion. The sampling distribution of $\bar{p}$, the point estimator of the population parameter p, is the basis for developing the test statistic.

When the null hypothesis is true as an equality, the expected value of $\bar{p}$ equals the hypothesized value p_0; that is, $E(\bar{p}) = p_0$. The standard error of $\bar{p}$ is given by

$$\sigma_{\bar{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

In Section 7.3, we said that if $np \geq 5$ and $n(1 - p) \geq 5$, the sampling distribution of $\bar{p}$ can be approximated by a normal distribution.[4] Under these conditions, which usually apply in practice, the quantity

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} \tag{7.12}$$

[4]In most applications involving hypothesis tests of a population proportion, sample sizes are large enough to use the normal approximation. The exact sampling distribution of $\bar{p}$ is discrete, with the probability for each value of $\bar{p}$ given by the binomial distribution. So hypothesis testing is a bit more complicated for small samples when the normal approximation cannot be used.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

has a standard normal probability distribution. With $\sigma_{\bar{p}} = \sqrt{p_0(1-p_0)/n}$, the standard normal random variable z is the test statistic used to conduct hypothesis tests about a population proportion.

Test Statistic for Hypothesis Tests about a Population Proportion

$$z = \frac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \tag{7.13}$$

DATA*file*
womengolf

We can now compute the test statistic for the Pine Creek hypothesis test. Suppose a random sample of 400 players was selected, and that 100 of the players were women. The proportion of women golfers in the sample is

$$\bar{p} = \frac{100}{400} = 0.25$$

Using equation (7.13), the value of the test statistic is

$$z = \frac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \frac{0.25 - 0.20}{\sqrt{\dfrac{0.20(1-0.20)}{400}}} = \frac{0.05}{0.02} = 2.50$$

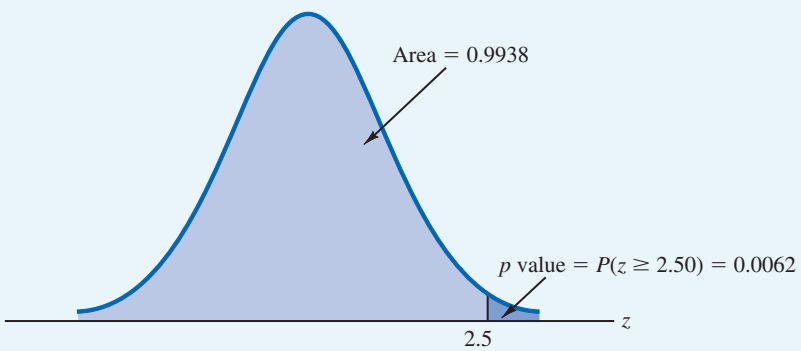
The Excel formula = NORM.S.DIST(z, TRUE) computes the area under the standard normal distribution curve that is less than or equal to the value z.

Because the Pine Creek hypothesis test is an upper-tail test, the p value is the probability of obtaining a value for the test statistic that is greater than or equal to $z = 2.50$; that is, it is the upper-tail area corresponding to $z \geq 2.50$ as displayed in Figure 7.22. The Excel formula $= 1 - NORM.S.DIST(2.5,TRUE)$ computes this upper-tail area of 0.0062.

Recall that the course manager specified a level of significance of $\alpha = 0.05$. A p value $= 0.0062 < 0.05$ gives sufficient statistical evidence to reject H_0 at the 0.05 level of significance. Thus, the test provides statistical support for the conclusion that the special promotion increased the proportion of women players at the Pine Creek golf course.

Using Excel Excel can be used to conduct one-tailed and two-tailed hypothesis tests about a population proportion using the p value approach. The procedure is similar to the approach used with Excel in conducting hypothesis tests about a population mean. The primary difference is that the test statistic is based on the sampling distribution of $\bar{x}$ for hypothesis tests about a population mean and on the sampling distribution of $\bar{p}$ for hypothesis tests about a population proportion. Thus, although different formulas are used to compute the test statistic and the p value needed to make the hypothesis testing decision, the logical process is identical.

Figure 7.22 Calculation of the *p* Value for the Pine Creek Hypothesis Test



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

MODEL file

womengolftest

The worksheet in Figure 7.23 can be used as a template for hypothesis tests about a population proportion whenever $np \geq 5$ *and* $n(1-p) \geq 5$. *Just enter the appropriate data in column A, adjust the ranges for the formulas in cells D3 and D5, enter the appropriate response in cell D4, and enter the hypothesized value in cell D8. The standard error, the test statistic, and the three* p *values will then appear. Depending on the form of the hypothesis test (lower-tail, upper-tail, or two-tailed), we can then choose the appropriate* p *value to make the rejection decision.*

We will illustrate the procedure by showing how Excel can be used to conduct the upper-tail hypothesis test for the Pine Creek golf course study. Refer to Figure 7.23 as we describe the tasks involved. The formula worksheet is on the left; the value worksheet is on the right.

The descriptive statistics needed are provided in cells D3, D5, and D6. Because the data are not numeric, Excel's COUNTA function, not the COUNT function, is used in cell D3 to determine the sample size. We entered *Female* in cell D4 to identify the response for which we wish to compute a proportion. The COUNTIF function is then used in cell D5 to determine the number of responses of the type identified in cell D4. The sample proportion is then computed in cell D6 by dividing the response count by the sample size.

The hypothesized value of the population proportion (*0.20*) is entered into cell D8. The standard error is obtained in cell D10 by entering the formula *=SQRT(D8*(1-D8)/D3)*. The formula *=(D6-D8)/D10* entered into cell D11 computes the test statistic z according to equation (7.13). To compute the p value for a lower-tail test, we enter the formula *= NORM.S.DIST(D11,TRUE)* into cell D13. The p value for an upper-tail test is tzhen computed in cell D14 as 1 minus the p value for the lower-tail test. Finally, the p value for a two-tailed test is computed in cell D15 as two times the minimum of the two one-tailed p values. The value worksheet shows that the three p values are as follows: p value (lower tail) = 0.9938, p value (upper tail) = 0.0062, and p value (two tail) = 0.0124.

The development of the worksheet is now complete. For the Pine Creek upper-tail hypothesis test, we reject the null hypothesis that the population proportion is 0.20 or less because the upper tail p value of 0.0062 is less than $\alpha = 0.05$. Indeed, with this p value we would reject the null hypothesis for any level of significance of 0.0062 or greater.

The procedure used to conduct a hypothesis test about a population proportion is similar to the procedure used to conduct a hypothesis test about a population mean. Although we illustrated how to conduct a hypothesis test about a population proportion only for an upper-tail test, similar procedures can be used for lower-tail and two-tailed tests. Table 7.8 provides a summary of the hypothesis tests about a population proportion for the case

Figure 7.23 Hypothesis Test for Pine Creek Golf Course

	A	B	C	D	E
1	Golfer		Hypothesis Test about a Population Mean:		
2	Female				
3	Male		Sample Size	=COUNTA(A2:A401)	
4	Female		Response of Interest	Female	Enter Female for Response of Interest
5	Male		Count for Response	=COUNTIF(A2:A401,D4)	
6	Male		Sample Mean	=D5/D3	
7	Female				
8	Male		Hypothesized Value	0.2	Enter 0.2 for Hypothesized Value
9	Male				
10	Female		Standard Error	=SQRT(D8*(1-D8)/D3)	
11	Male		Test Statistic z	=(D6-D8)/D10	
12	Male				
13	Male		*p*-value (Lower Tail)	=NORM.S.DIST(D11,TRUE)	
14	Male		*p*-value (Upper Tail)	=1-D13	
15	Male		*p*-value (Two Tail)	=2*(MIN(D13,D14))	
16	Female				
400	Male				
401	Male				

	A	B	C	D
1	Golfer		Hypothesis Test about a Population Mean:	
2	Female			
3	Male		Sample Size	400
4	Female		Response of Interest	Female
5	Male		Count for Response	100
6	Male		Sample Mean	0.250
7	Female			
8	Male		Hypothesized Value	0.2
9	Male			
10	Female		Standard Error	0.0200
11	Male		Test Statistic z	2.5000
12	Male			
13	Male		*p*-value (Lower Tail)	0.9938
14	Male		*p*-value (Upper Tail)	0.0062
15	Male		*p*-value (Two Tail)	0.0124
16	Female			
400	Male			
401	Male			
402				

Note: Rows 17–399 are hidden.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 7.8 Summary of Hypothesis Tests About a Population Proportion

	Lower-Tail Test	Upper-Tail Test	Two-Tailed Test
Hypotheses	$H_0: p \geq p_0$ $H_a: p < p_0$	$H_0: p \leq p_0$ $H_a: p > p_0$	$H_0: p = p_0$ $H_a: p \neq p_0$
Test Statistic	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$
***p* Value**	=NORM.S.DIST (z, TRUE)	=1 − NORM.S.DIST (z, TRUE)	2*MIN(NORM.S.DIST(z, TRUE), 1 − NORM.S.DIST(z, TRUE))

that $np \geq 5$ and $n(1-p) \geq 5$ (and thus the normal probability distribution can be used to approximate the sampling distribution of $\bar{p}$).

Notes + Comments

1. We have shown how to use *p* values. The smaller the *p* value, the stronger the evidence in the sample data against H_0 and the stronger the evidence in favor of H_a. Here are guidelines that some statisticians suggest for interpreting small *p* values:
   - Less than 0.01—overwhelming evidence to conclude that H_a is true
   - Between 0.01 and 0.05—strong evidence to conclude that H_a is true
   - Between 0.05 and 0.10—weak evidence to conclude that H_a is true
   - Greater than 0.10—insufficient evidence to conclude that H_a is true
2. The procedures for testing hypotheses about the mean that are discussed in this chapter are reliable unless the sample size is small and the population is highly skewed or contains outliers. In these cases, a nonparametric approach such as the sign test can be used. Under these conditions the results of nonparametric tests are more reliable than the hypothesis testing procedures discussed in this chapter. However, this increased reliability comes with a cost; if the sample is large or the population is relatively normally distributed, a nonparametric approach will also reject false null hypotheses less frequently.
3. We have discussed only procedures for testing hypotheses about the mean or proportion of a single population. There are many statistical procedures for testing hypotheses about multiple means or proportions. There are also many statistical procedures for testing hypotheses about parameters other than the population mean or the population proportion.

7.6 Big Data, Statistical Inference, and Practical Significance

As stated earlier in this chapter, the purpose of statistical inference is to use sample data to quickly and inexpensively gain insight into some characteristic of a population. Therefore, it is important that we can expect the sample to look like, or be representative of, the population that is being investigated. In practice, individual samples always, to varying degrees, fail to be perfectly representative of the population of interest. There are two general reasons a sample may fail to be representative of the population of interest: sampling error and nonsampling error.

Sampling Error

One reason a sample may fail to represent the population of interest is sampling error, or deviation of the sample from the population that results from random sampling. If repeated independent random samples of the same size are collected from the population of interest using a probability sampling techniques, on average the samples will be representative of the population from which the samples have been taken. This is the justification for collecting sample data randomly. However, the random collection of sample data does not ensure that any single sample will be perfectly representative of the population from which the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

sample has been taken; when collecting a sample randomly, the data in the sample cannot be expected to be perfectly representative of the population from which it has been taken. Sampling error is unavoidable when collecting a random sample; this is a risk we must accept when we chose to collect a random sample rather than incur the costs associated with taking a census of the population.

As expressed by equations (7.2) and (7.5), the standard errors of the sampling distributions of the sample mean $\overline{x}$ and the sample proportion of $\overline{p}$ reflect the potential for sampling error when using sample data to estimate the population mean μ and the population proportion p, respectively. As the sample size n increases, the potential impact of extreme values on the statistic decreases, so there is less variation in the potential values of the statistic produced by the sample and the standard errors of these sampling distributions decrease. Because these standard errors reflect the potential for sampling error when using sample data to estimate the population mean μ and the population proportion p, we see that for an extremely large sample there may be little potential for sampling error.

Nonsampling Error

Although the standard error of a sampling distribution decreases as the sample size n increases, this does not mean that we can conclude that an extremely large sample will always provide reliable information about the population of interest; this is because sampling error is not the sole reason a sample may fail to represent the target population. Deviations of the sample from the population that occur for reasons other than random sampling are referred to as **nonsampling error**. Nonsampling error can occur for a variety of reasons.

Consider the online news service PenningtonDailyTimes.com (PDT). Because PDT's primary source of revenue is the sale of advertising, the news service is intent on collecting sample data on the behavior of visitors to its web site in order to support its advertising sales. Prospective advertisers are willing to pay a premium to advertise on web sites that have long visit times, so PDT's management is keenly interested in the amount of time customers spend during their visits to PDT's web site. Advertisers are also concerned with how frequently visitors to a web site click on any of the ads featured on the web site, so PDT is also interested in whether visitors to its web site clicked on any of the ads featured on PenningtonDailyTimes.com.

From whom should PDT collect its data? Should it collect data on current visits to PenningtonDailyTimes.com? Should it attempt to attract new visitors and collect data on these visits? If so, should it measure the time spent at its web site by visitors it has attracted from competitors' web sites or visitors who do not routinely visit online news sites? The answers to these questions depend on PDT's research objectives. Is the company attempting to evaluate its current market, assess the potential of customers it can attract from competitors, or explore the potential of an entirely new market such as individuals who do not routinely obtain their news from online news services? If the research objective and the population from which the sample is to be drawn are not aligned, the data that PDT collects will not help the company accomplish its research objective. This type of error is referred to as a **coverage error**.

Nonsampling error can occur in a sample or a census.

Even when the sample is taken from the appropriate population, nonsampling error can occur when segments of the target population are systematically underrepresented or overrepresented in the sample. This may occur because the study design is flawed or because some segments of the population are either more likely or less likely to respond. Suppose PDT implements a pop-up questionnaire that opens when a visitor leaves PenningtonDailyTimes.com. Visitors to PenningtonDailyTimes.com who have installed pop-up blockers will be likely underrepresented, and visitors to PenningtonDailyTimes.com who have not installed pop-up blockers will likely be overrepresented. If the behavior of PenningtonDailyTimes.com visitors who have installed pop-up blockers differs from the behaviors of PenningtonDailyTimes.com visitors who have not installed pop-up blockers,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

attempting to draw conclusions from this sample about how all visitors to the PDT web site behave may be misleading. This type of error is referred to as a **nonresponse error**.

Another potential source of nonsampling error is incorrect measurement of the characteristic of interest. This type of error is referred to as a **measurement error**. For example, if PDT asks questions that are ambiguous or difficult for respondents to understand, the responses may not accurately reflect how the respondents intended to respond. For example, respondents may be unsure how to respond if PDT asks "Are the news stories on PenningtonDailyTimes.com compelling and accurate?". How should a visitor respond if they feel the news stories on PenningtonDailyTimes.com are compelling but erroneous? What response is appropriate if the respondent feels the news stories on PenningtonDailyTimes.com are accurate but dull? A similar issue can arise if a question is asked in a biased or leading way. If PDT asks "Many readers find the news stories on PenningtonDailyTimes.com to be compelling and accurate. Do you find the news stories on PenningtonDailyTimes.com to be compelling and accurate?", the qualifying statement PDT makes prior to the actual question will likely result in a bias toward positive responses. Incorrect measurement of the characteristic of interest can also occur when respondents provide incorrect answers; this may be due to a respondent's poor recall or unwillingness to respond honestly.

Errors that are introduced by interviewers or during the recording and preparation of the data are other types of nonsampling error. These types of error are referred to as interviewer errors and processing errors, respectively.

Nonsampling error can introduce bias into the estimates produced using the sample, and this bias can mislead decision makers who use the sample data in their decision-making processes. No matter how small or large the sample, we must contend with this limitation of sampling whenever we use sample data to gain insight into a population of interest. Although sampling error decreases as the size of the sample increases, an extremely large sample can still suffer from nonsampling error and fail to be representative of the population of interest. When sampling, care must be taken to ensure that we minimize the introduction of nonsampling error into the data collection process. This can be done by carrying out the following steps:

- Carefully define the target population before collecting sample data, and subsequently design the data collection procedure so that a probability sample is drawn from this target population.
- Carefully design the data collection process and train the data collectors.
- Pretest the data collection procedure to identify and correct for potential sources of nonsampling error prior to final data collection.
- Use stratified random sampling when population-level information about an important qualitative variable is available to ensure that the sample is representative of the population with respect to that qualitative characteristic.
- Use cluster sampling when the population can be divided into heterogeneous subgroups or clusters.
- Use systematic sampling when population-level information about an important quantitative variable is available to ensure that the sample is representative of the population with respect to that quantitative characteristic.

Finally, recognize that every random sample (even an extremely large random sample) will suffer from some degree of sampling error, and eliminating all potential sources of nonsampling error may be impractical. Understanding these limitations of sampling will enable us to be more realistic and pragmatic when interpreting sample data and using sample data to draw conclusions about the target population.

Big Data

Recent estimates state that approximately 2.5 quintillion bytes of data are created worldwide each day. This represents a dramatic increase from the estimated 100 gigabytes (GB) of data generated worldwide per day in 1992, the 100 GB of data generated worldwide per hour in 1997, and the 100 GB of data generated worldwide per second in 2002. Every

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

minute, there is an average of 216,000 Instagram posts, 204,000,000 e-mails sent, 12 hours of footage uploaded to YouTube, and 277,000 tweets posted on Twitter. Without question, the amount of data that is now generated is overwhelming, and this trend is certainly expected to continue.

In each of these cases the data sets that are generated are so large or complex that current data processing capacity and/or analytic methods are not adequate for analyzing the data. Thus, each is an example of **big data**. There are myriad other sources of big data. Sensors and mobile devices transmit enormous amounts of data. Internet activities, digital processes, and social media interactions also produce vast quantities of data.

The amount of data has increased so rapidly that our vocabulary for describing a data set by its size must expand. A few years ago, a petabyte of data seemed almost unimaginably large, but we now routinely describe data in terms of yottabytes. Table 7.9 summarizes terminology for describing the size of data sets.

Understanding What Big Data Is

The processes that generate big data can be described by four attributes or dimensions that are referred to as the four V's:

- **Volume**—the amount of data generated
- **Variety**—the diversity in types and structures of data generated
- **Veracity**—the reliability of the data generated
- **Velocity**—the speed at which the data are generated

A high degree of any of these attributes individually is sufficient to generate big data, and when they occur at high levels simultaneously the resulting amount of data can be overwhelmingly large. Technological advances and improvements in electronic (and often automated) data collection make it easy to collect millions, or even billions, of observations in a relatively short time. Businesses are collecting greater volumes of an increasing variety of data at a higher velocity than ever.

To understand the challenges presented by big data, we consider its structural dimensions. Big data can be **tall data**; a data set that has so many observations that traditional statistical inference has little meaning. For example, producers of consumer goods collect information on the sentiment expressed in millions of social media posts each day to better understand consumer perceptions of their products. Such data consist of the sentiment expressed (the variable) in millions (or over time, even billions) of social media posts (the observations). Big data can also be **wide data**; a data set that has so many variables that simultaneous consideration of all variables is infeasible. For example, a high-resolution image can comprise millions or billions of pixels. The data used by facial recognition algorithms consider each pixel in an image when comparing an image to other images in an attempt to find a match. Thus, these algorithms make use of the characteristics of millions

Table 7.9 Terminology for Describing the Size of Data Sets

Number of Bytes	Metric	Name
1000^1	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	yottabyte

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

or billions of pixels (the variables) for relatively few high-resolution images (the observations). Of course, big data can be both tall and wide, and the resulting data set can again be overwhelmingly large.

Statistics are useful tools for understanding the information embedded in a big data set, but we must be careful when using statistics to analyze big data. It is important that we understand the limitations of statistics when applied to big data and we temper our interpretations accordingly. Because tall data are the most common form of big data used in business, we focus on this structure in the discussions throughout the remainder of this section.

Big Data and Sampling Error

A sample of one million or more visitors might seem unrealistic, but keep in mind that according to techjury.net, amazon.com has an average 197 million visitors per month.

Let's revisit the data collection problem of online news service PenningtonDailyTimes.com (PDT). Because PDT's primary source of revenue is the sale of advertising, PDT's management is interested in the amount of time customers spend during their visits to PDT's web site. From historical data, PDT has estimated that the standard deviation of the time spent by individual customers when they visit PDT's web site is $s = 20$ seconds. Table 7.10 shows how the standard error of the sampling distribution of the sample mean time spent by individual customers when they visit PDT's web site decreases as the sample size increases.

PDT also wants to collect information from its sample respondents on whether a visitor to its web site clicked on any of the ads featured on the web site. From its historical data, PDT knows that 51% of past visitors to its web site clicked on an ad featured on the web site, so it will use this value as $\bar{p}$ to estimate the standard error. Table 7.11 shows how the standard error of the sampling distribution of the proportion of the sample that clicked on any of the ads featured on PenningtonDailyTimes.com decreases as the sample size increases.

The PDT example illustrates the general relationship between standard errors and the sample size. We see in Table 7.10 that the standard error of the sample mean decreases as the sample size increases. For a sample of $n = 10$, the standard error of the sample mean is 6.32456; when we increase the sample size to $n = 100{,}000$, the standard error of the sample mean decreases to 0.06325; and at a sample size of $n = 1{,}000{,}000{,}000$, the standard error of the sample mean decreases to only 0.00063. In Table 7.11 we see that the standard error of the sample proportion also decreases as the sample size increases. For a sample of $n = 10$, the standard error of the sample proportion is 0.15808; when we increase the sample size to $n = 100{,}000$, the standard error of the sample proportion decreases to 0.00158; and at a sample size of $n = 1{,}000{,}000{,}000$, the standard error of the sample mean decreases to only 0.00002. In both Table 7.10 and Table 7.11, the standard error when $n = 1{,}000{,}000{,}000$ is *one ten-thousandth of the standard error when* $n = 10$.

Table 7.10 Standard Error of the Sample Mean $\bar{x}$ When $s = 20$ at Various Sample Sizes n

Sample Size n	Standard Error $s_{\bar{x}} = s/\sqrt{n}$
10	6.32456
100	2.00000
1,000	0.63246
10,000	0.20000
100,000	0.06325
1,000,000	0.02000
10,000,000	0.00632
100,000,000	0.00200
1,000,000,000	0.00063

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 7.11 Standard Error of the Sample Proportion $\bar{p}$ When $p = 0.51$ at Various Sample Sizes n

Sample Size n	Standard Error $\sigma_{\bar{p}} = \sqrt{\bar{p}(1-\bar{p})/n}$
10	0.15808
100	0.04999
1,000	0.01581
10,000	0.00500
100,000	0.00158
1,000,000	0.00050
10,000,000	0.00016
100,000,000	0.00005
1,000,000,000	0.00002

Big Data and the Precision of Confidence Intervals

We have seen that confidence intervals are powerful tools for making inferences about population parameters, but the validity of any interval estimate depends on the quality of the data used to develop the interval estimate. No matter how large the sample is, if the sample is not representative of the population of interest, the confidence interval cannot provide useful information about the population parameter of interest. In these circumstances, statistical inference can be misleading.

A review of equations (7.7) and (7.10) shows that confidence intervals for the population mean μ and population proportion p become more narrow as the size of the sample increases. Therefore, the potential sampling error also decreases as the sample size increases. To illustrate the rate at which interval estimates narrow for a given confidence level, we again consider the PenningtonDailyTimes.com (PDT) example.

Recall that PDT's primary source of revenue is the sale of advertising, and prospective advertisers are willing to pay a premium to advertise on web sites that have long visit times. Suppose PDT's management wants to develop a 95% confidence interval estimate of the mean amount of time customers spend during their visits to PDT's web site. Table 7.12 shows how the margin of error at the 95% confidence level decreases as the sample size increases when $s = 20$.

Suppose that in addition to estimating the population mean amount of time customers spend during their visits to PDT's web site, PDT would like to develop a 95% confidence interval estimate of the proportion of its web site visitors that click on an ad. Table 7.13 shows how the margin of error for a 95% confidence interval estimate of the population proportion decreases as the sample size increases when the sample proportion is $\bar{p} = 0.51$.

The PDT example illustrates the relationship between the precision of interval estimates and the sample size. We see in Tables 7.12 and 7.13 that at a given confidence level, the margins of error decrease as the sample sizes increase. As a result, if the sample mean time spent by customers when they visit PDT's web site is 84.1 seconds, the 95% confidence interval estimate of the population mean time spent by customers when they visit PDT's web site decreases from (69.79286, 98.40714) for a sample of $n = 10$ to (83.97604, 84.22396) for a sample of $n = 100{,}000$ to (84.09876, 84.10124) for a sample of $n = 1{,}000{,}000{,}000$. Similarly, if the sample proportion of its web site visitors who clicked on an ad is 0.51, the 95% confidence interval estimate of the population proportion of its web site visitors who clicked on an ad decreases from (0.20016, 0.81984) for a sample of $n = 10$ to (0.50690, 0.51310) for a sample of $n = 100{,}000$ to (0.50997, 0.51003) for a

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 7.12 Margin of Error for Interval Estimates of the Population Mean at the 95% Confidence Level for Various Sample Sizes n

Sample Size n	Margin of Error $t_{\alpha/2}s_{\bar{x}}$
10	14.30714
100	3.96843
1,000	1.24109
10,000	0.39204
100,000	0.12396
1,000,000	0.03920
10,000,000	0.01240
100,000,000	0.00392
1,000,000,000	0.00124

Table 7.13 Margin of Error for Interval Estimates of the Population Proportion at the 95% Confidence Level for Various Sample Sizes n

Sample Size n	Margin of Error $z_{\alpha/2}\sigma_{\bar{p}}$
10	0.30984
100	0.09798
1,000	0.03098
10,000	0.00980
100,000	0.00310
1,000,000	0.00098
10,000,000	0.00031
100,000,000	0.00010
1,000,000,000	0.00003

sample of $n = 1{,}000{,}000{,}000$. In both instances, as the sample size becomes extremely large, the margin of error becomes extremely small and the resulting confidence intervals become extremely narrow.

Implications of Big Data for Confidence Intervals

Last year, the mean time spent by all visitors to PenningtonDailyTimes.com was 84 seconds. Suppose that PDT wants to assess whether the population mean time has changed since last year. PDT now collects a new sample of 1,000,000 visitors to its web site and calculates the sample mean time spent by these visitors to the PDT web site to be $\bar{x} = 84.1$ seconds. The estimated population standard deviation is $s = 20$ seconds, so the standard error is $s_{\bar{x}} = s/\sqrt{n} = 0.02000$. Furthermore, the sample is sufficiently large to ensure that the sampling distribution of the sample mean will be normally distributed. Thus, the 95% confidence interval estimate of the population mean is

$$\bar{x} \pm t_{\alpha/2}s_{\bar{x}} = 84.1 \pm 0.0392 = (84.06080, 84.13920)$$

What could PDT conclude from these results? There are three possible reasons that PDT's sample mean of 84.1 seconds differs from last year's population mean of 84 seconds: (1) sampling error, (2) nonsampling error, or (3) the population mean has changed

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

since last year. The 95% confidence interval estimate of the population mean does not include the value for the mean time spent by all visitors to the PDT web site for last year (84 seconds), suggesting that the difference between PDT's sample mean for the new sample (84.1 seconds) and the mean from last year (84 seconds) is not likely to be exclusively a consequence of sampling error. Nonsampling error is a possible explanation and should be investigated as the results of statistical inference become less reliable as nonsampling error is introduced into the sample data. If PDT determines that it introduced little or no nonsampling error into its sample data, the only remaining plausible explanation for a difference of this magnitude is that the population mean has changed since last year.

If PDT concludes that the sample has provided reliable evidence and the population mean has changed since last year, management must still consider the potential impact of the difference between the sample mean and the mean from last year. If a 0.1 second difference in the time spent by visitors to PenningtonDailyTimes.com has a consequential effect on what PDT can charge for advertising on its site, this result could have practical business implications for PDT. Otherwise, there may be no **practical significance** of the 0.1 second difference in the time spent by visitors to PenningtonDailyTimes.com.

Confidence intervals are extremely useful, but as with any other statistical tool, they are only effective when properly applied. Because interval estimates become increasingly precise as the sample size increases, extremely large samples will yield extremely precise estimates. However, no interval estimate, no matter how precise, will accurately reflect the parameter being estimated unless the sample is relatively free of nonsampling error. Therefore, when using interval estimation, it is always important to carefully consider whether a random sample of the population of interest has been taken.

Big Data, Hypothesis Testing, and *p* Values

We have seen that interval estimates of the population mean μ and the population proportion p narrow as the sample size increases. This occurs because the standard error of the associated sampling distributions decrease as the sample size increases. Now consider the relationship between interval estimation and hypothesis testing that we discussed earlier in this chapter. If we construct a $100(1-\alpha)\%$ interval estimate for the population mean, we reject $H_0: \mu = \mu_0$ if the $100(1-\alpha)\%$ interval estimate does not contain μ_0. Thus, for a given level of confidence, as the sample size increases we will reject $H_0: \mu = \mu_0$ for increasingly smaller differences between the sample mean $\bar{x}$ and the hypothesized population mean μ_0. We can see that when the sample size n is very large, almost any difference between the sample mean $\bar{x}$ and the hypothesized population mean μ_0 results in rejection of the null hypothesis.

In this section, we will elaborate how big data affects hypothesis testing and the magnitude of p values. Specifically, we will examine how rapidly the p value associated with a given difference between a point estimate and a hypothesized value of a parameter decreases as the sample size increases.

Let us again consider the online news service PenningtonDailyTimes.com (PDT). Recall that PDT's primary source of revenue is the sale of advertising, and prospective advertisers are willing to pay a premium to advertise on web sites that have long visit times. To promote its news service, PDT's management wants to promise potential advertisers that the mean time spent by customers when they visit PenningtonDailyTimes.com is greater than last year, that is, more than 84 seconds. PDT therefore decides to collect a sample tracking the amount of time spent by individual customers when they visit PDT's web site in order to test its null hypothesis $H_0: \mu \leq 84$.

For a sample mean of 84.1 seconds and a sample standard deviation of $s = 20$ seconds, Table 7.14 provides the values of the test statistic t and the p values for the test of the null hypothesis $H_0: \mu \leq 84$. The p value for this hypothesis test is essentially 0 for all samples in Table 7.14 with at least $n = 1{,}000{,}000$.

PDT's management also wants to promise potential advertisers that the proportion of its web site visitors who click on an ad this year exceeds the proportion of its web site visitors who clicked on an ad last year, which was 0.50. PDT collects information from its sample

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 7.14 Values of the Test Statistic t and the p Values for the Test of the Null Hypothesis H_0: $\mu \leq 84$ for Various Sample Sizes When Sample Mean $\bar{x} = 84.1$ Seconds and Sample Standard Deviation $s = 20$

Sample Size n	t	p Value
10	0.01581	0.49386
100	0.05000	0.48011
1,000	0.15811	0.43720
10,000	0.50000	0.30854
100,000	1.58114	0.05692
1,000,000	5.00000	2.87E-07
10,000,000	15.81139	1.30E-56
100,000,000	50.00000	0.00E+00
1,000,000,000	158.11388	0.00E+00

on whether the visitor to its web site clicked on any of the ads featured on the web site, and it wants to use these data to test its null hypothesis H_0: $p \leq 0.50$.

For a sample proportion of 0.51, Table 7.15 provides the values of the test statistic z and the p values for the test of the null hypothesis H_0: $p \leq 0.50$. The p value for this hypothesis test is essentially 0 for all samples in Table 7.14 with at least $n = 100{,}000$.

We see in Tables 7.14 and 7.15 that the p value associated with a given difference between a point estimate and a hypothesized value of a parameter decreases as the sample size increases. As a result, if the sample mean time spent by customers when they visit PDT's web site is 84.1 seconds, PDT's null hypothesis H_o: $\mu \leq 84$ is not rejected at $\alpha = 0.01$ for samples with $n \leq 100{,}000$, and is rejected at $\alpha = 0.01$ for samples with $n \geq 1{,}000{,}000$. Similarly, if the sample proportion of visitors to its web site clicked on an ad featured on the web site is 0.51, PDT's null hypothesis H_0: $p \leq 0.50$ is not rejected at $\alpha = 0.01$ for samples with $n \leq 10{,}000$, and is rejected at $\alpha = 0.01$ for samples with $n \geq 100{,}000$. In both instances, as the sample size becomes extremely large the p value associated with the given difference between a point estimate and the hypothesized value of the parameter becomes extremely small.

Table 7.15 Values of the Test Statistic z and the p Values for the Test of the Null Hypothesis H_0: $p \leq .50$ for Various Sample Sizes When Sample Proportion $\bar{p} = 0.51$

Sample Size n	z	p Value
10	0.06325	0.47479
100	0.20000	0.42074
1,000	0.63246	0.26354
10,000	2.00000	0.02275
100,000	6.32456	1.27E-10
1,000,000	20.00000	0.00E+00
10,000,000	63.24555	0.00E+00
100,000,000	200.00000	0.00E+00
1,000,000,000	632.45553	0.00E+00

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Implications of Big Data in Hypothesis Testing

Suppose PDT collects a sample of 1,000,000 visitors to its web site and uses these data to test its null hypotheses H_0: $\mu \leq 84$ and H_0: $p \leq 0.50$ at the 0.05 level of significance. As the sixth rows of Tables 7.14 and 7.15 show, respectively, the null hypothesis is rejected in both tests. As a result, PDT can promise potential advertisers that the mean time spent by individual customers who visit PDT's web site exceeds 84 seconds and the proportion of individual visitors to its web site who click on an ad exceeds 0.50. These results suggest that for each of these hypothesis tests, the difference between the point estimate and the hypothesized value of the parameter being tested is not likely solely a consequence of sampling error. However, the results of any hypothesis test, no matter the sample size, are only reliable if the sample is relatively free of nonsampling error. If nonsampling error is introduced in the data collection process, the likelihood of making a Type I or Type II error may be higher than if the sample data are free of nonsampling error. Therefore, when testing a hypothesis, it is always important to think carefully about whether a random sample of the population of interest has been taken.

If PDT determines that it has introduced little or no nonsampling error into its sample data, the only remaining plausible explanation for these results is that these null hypotheses are false. At this point, PDT and the companies that advertise on PenningtonDailyTimes.com should also consider whether these statistically significant differences between the point estimates and the hypothesized values of the parameters being tested are of practical significance. Although a 0.1 second increase in the mean time spent by customers when they visit PDT's web site is statistically significant, it may not be meaningful to companies that might advertise on PenningtonDailyTimes.com. Similarly, although an increase of 0.01 in the proportion of visitors to its web site that click on an ad is statistically significant, it may not be meaningful to companies that might advertise on PenningtonDailyTimes.com. One has to determine whether these statistically significant differences have meaningful implications for ensuing business decisions of PDT and its advertisers.

Ultimately, no business decision should be based solely on statistical inference. Practical significance should always be considered in conjunction with statistical significance. This is particularly important when the hypothesis test is based on an extremely large sample because even an extremely small difference between the point estimate and the hypothesized value of the parameter being tested will be statistically significant. When done properly, statistical inference provides evidence that should be considered in combination with information collected from other sources to make the most informed decision possible.

Notes + Comments

1. Nonsampling error can occur when either a probability sampling technique or a nonprobability sampling technique is used. However, nonprobability sampling techniques such as convenience sampling and judgment sampling often introduce nonsampling error into sample data because of the manner in which sample data are collected. Therefore, probability sampling techniques are preferred over nonprobability sampling techniques.
2. When taking an extremely large sample, it is conceivable that the sample size is at least 5% of the population size; that is, $n/N \geq 0.05$. Under these conditions, it is necessary to use the finite population correction factor when calculating the standard error of the sampling distribution to be used in confidence intervals and hypothesis testing.

Summary

In this chapter, we presented the concepts of sampling and sampling distributions. We demonstrated how a simple random sample can be selected from a finite population and how a random sample can be selected from an infinite population. The data collected from such samples can be used to develop point estimates of population parameters. Different

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

samples provide different values for the point estimators; therefore, point estimators such as $\bar{x}$ and $\bar{p}$ are random variables. The probability distribution of such a random variable is called a sampling distribution. In particular, we described in detail the sampling distributions of the sample mean $\bar{x}$ and the sample proportion $\bar{p}$. In considering the characteristics of the sampling distributions of $\bar{x}$ and $\bar{p}$, we stated that $E(\bar{x}) = \mu$ and $E(\bar{p}) = p$. After developing the standard deviation or standard error formulas for these estimators, we described the conditions necessary for the sampling distributions of $\bar{x}$ and $\bar{p}$ to follow a normal distribution.

In Section 7.4, we presented methods for developing interval estimates of a population mean and a population proportion. A point estimator may or may not provide a good estimate of a population parameter. The use of an interval estimate provides a measure of the precision of an estimate. Both the interval estimate of the population mean and the population proportion take the form: point estimate $\pm$ margin of error.

We presented the interval estimation procedure for a population mean for the practical case in which the population standard deviation is unknown. The interval estimation procedure uses the sample standard deviation s and the t distribution. The quality of the interval estimate obtained depends on the distribution of the population and the sample size. In a normally distributed population, the interval estimates will be exact in both cases, even for small sample sizes. If the population is not normally distributed, the interval estimates obtained will be approximate. Larger sample sizes provide better approximations, but the more highly skewed the population is, the larger the sample size needs to be to obtain a good approximation.

The general form of the interval estimate for a population proportion is $\bar{p} \pm$ margin of error. In practice, the sample sizes used for interval estimates of a population proportion are generally large. Thus, the interval estimation procedure for a population proportion is based on the standard normal distribution.

In Section 7.5, we presented methods for hypothesis testing, a statistical procedure that uses sample data to determine whether or not a statement about the value of a population parameter should be rejected. The hypotheses are two competing statements about a population parameter. One statement is called the null hypothesis (H_0), and the other is called the alternative hypothesis (H_a). We provided guidelines for developing hypotheses for situations frequently encountered in practice.

In the hypothesis-testing procedure for the population mean, the sample standard deviation s is used to estimate σ and the hypothesis test is based on the t distribution. The quality of results depends on both the form of the population distribution and the sample size; if the population is not normally distributed, larger sample sizes are needed. General guidelines about the sample size were provided in Section 7.5. In the case of hypothesis tests about a population proportion, the hypothesis-testing procedure uses a test statistic based on the standard normal distribution.

We also reviewed how the value of the test statistic can be used to compute a p value—a probability that is used to determine whether the null hypothesis should be rejected. If the p value is less than or equal to the level of significance α, the null hypothesis can be rejected.

In Section 7.6, we discussed the concept of big data and its implications for statistical inference. We considered sampling and nonsampling error; the implications of big data on standard errors, confidence intervals, and hypothesis testing for the mean and the proportion; and the importance of considering both statistical significance and practical significance.

Glossary

Alternative hypothesis The hypothesis concluded to be true if the null hypothesis is rejected.

Big data Any set of data that is too large or too complex to be handled by standard data processing techniques and typical desktop software.

Census Collection of data from every element in the population of interest.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Central limit theorem A theorem stating that when enough independent random variables are added, the resulting sum is a normally distributed random variable. This result allows one to use the normal probability distribution to approximate the sampling distributions of the sample mean and sample proportion for sufficiently large sample sizes.

Confidence coefficient The confidence level expressed as a decimal value. For example, 0.95 is the confidence coefficient for a 95% confidence level.

Confidence interval Another name for an interval estimate.

Confidence level The confidence associated with an interval estimate. For example, if an interval estimation procedure provides intervals such that 95% of the intervals formed using the procedure will include the population parameter, the interval estimate is said to be constructed at the 95% confidence level.

Coverage error Nonsampling error that results when the research objective and the population from which the sample is to be drawn are not aligned.

Degrees of freedom A parameter of the t distribution. When the t distribution is used in the computation of an interval estimate of a population mean, the appropriate t distribution has $n - 1$ degrees of freedom, where n is the size of the sample.

Finite population correction factor The term $\sqrt{(N - n)/(N - 1)}$ that is used in the formulas for computing the (estimated) standard error for the sample mean and sample proportion whenever a finite population, rather than an infinite population, is being sampled. The generally accepted rule of thumb is to ignore the finite population correction factor whenever $n/N \leq 0.05$.

Frame A listing of the elements from which the sample will be selected.

Hypothesis testing The process of making a conjecture about the value of a population parameter, collecting sample data that can be used to assess this conjecture, measuring the strength of the evidence against the conjecture that is provided by the sample, and using these results to draw a conclusion about the conjecture.

Interval estimate An estimate of a population parameter that provides an interval believed to contain the value of the parameter. For the interval estimates in this chapter, it has the form: point estimate $\pm$ margin of error.

Interval estimation The process of using sample data to calculate a range of values that is believed to include the unknown value of a population parameter.

Level of significance The probability that the interval estimation procedure will generate an interval that does not contain the value of parameter being estimated; also the probability of making a Type I error when the null hypothesis is true as an equality.

Margin of error The $\pm$ value added to and subtracted from a point estimate in order to develop an interval estimate of a population parameter.

Measurement error Nonsampling error that results when the values of one or more variables are not accurately measured or recorded.

Nonresponse error Nonsampling error that results when some segments of the population are more likely or less likely to respond to the survey mechanism.

Nonsampling error Any difference between the value of a sample statistic (such as the sample mean, sample standard deviation, or sample proportion) and the value of the corresponding population parameter (population mean, population standard deviation, or population proportion) that are not the result of sampling error. These include but are not limited to coverage error, nonresponse error, measurement error, interviewer error, and processing error.

Null hypothesis The hypothesis tentatively assumed to be true in the hypothesis testing procedure.

One-tailed test A hypothesis test in which rejection of the null hypothesis occurs for values of the test statistic in one tail of its sampling distribution.

***p* value** The probability, assuming that H_0 is true, of obtaining a random sample of size n that results in a test statistic at least as extreme as the one observed in the current sample. For a lower-tail test, the p value is the probability of obtaining a value for the test statistic as small as or smaller than that provided by the sample. For an upper-tail test, the p value is the probability of obtaining a value for the test statistic as large as or larger than that provided by

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the sample. For a two-tailed test, the p value is the probability of obtaining a value for the test statistic at least as unlikely as or more unlikely than that provided by the sample.

Parameter A measurable factor that defines a characteristic of a population, process, or system, such as a population mean μ, a population standard deviation σ, or a population proportion p.

Point estimate The value of a point estimator used in a particular instance as an estimate of a population parameter.

Point estimator The sample statistic, such as $\bar{x}$, s, or $\bar{p}$, that provides the point estimate of the population parameter.

Practical significance The real-world impact the result of statistical inference will have on business decisions.

Random sample A random sample from an infinite population is a sample selected such that the following conditions are satisfied: (1) Each element selected comes from the same population and (2) each element is selected independently.

Random variable A quantity whose values are not known with certainty.

Sample statistic A characteristic of sample data, such as a sample mean $\bar{x}$, a sample standard deviation s, or a sample proportion $\bar{p}$. The value of the sample statistic is used to estimate the value of the corresponding population parameter.

Sampled population The population from which the sample is drawn.

Sampling distribution A probability distribution consisting of all possible values of a sample statistic.

Sampling error The difference between the value of a sample statistic (such as the sample mean, sample standard deviation, or sample proportion) and the value of the corresponding population parameter (population mean, population standard deviation, or population proportion) that occurs because a random sample is used to estimate the population parameter.

Simple random sample A simple random sample of size n from a finite population of size N is a sample selected such that each possible sample of size n has the same probability of being selected.

Standard error The standard deviation of a point estimator.

Standard normal distribution A normal distribution with a mean of zero and standard deviation of one.

Statistical inference The process of making estimates and drawing conclusions about one or more characteristics of a population (the value of one or more parameters) through the analysis of sample data drawn from the population.

***t* distribution** A family of probability distributions that can be used to develop an interval estimate of a population mean whenever the population standard deviation s is unknown and is estimated by the sample standard deviation s.

Tall data A data set that has so many observations that traditional statistical inference has little meaning.

Target population The population for which statistical inferences such as point estimates are made. It is important for the target population to correspond as closely as possible to the sampled population.

Test statistic A statistic whose value helps determine whether a null hypothesis should be rejected.

Two-tailed test A hypothesis test in which rejection of the null hypothesis occurs for values of the test statistic in either tail of its sampling distribution.

Type I error The error of rejecting H_0 when it is true.

Type II error The error of accepting H_0 when it is false.

Unbiased A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter it estimates.

Variety The diversity in types and structures of data generated.

Velocity The speed at which the data are generated.

Veracity The reliability of the data generated.

Volume The amount of data generated.

Wide data A data set that has so many variables that simultaneous consideration of all variables is infeasible.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Problems

DATA*file* americanleague

1. **Randomly Sampling American League Teams.** The American League consists of 15 baseball teams. Suppose a sample of 5 teams is to be selected to conduct player interviews. The following table lists the 15 teams and the random numbers assigned by Excel's RAND function. Use these random numbers to select a sample of size 5. LO 1

Team	Random Number	Team	Random Number
New York	0.178624	Boston	0.290197
Baltimore	0.578370	Tampa Bay	0.867778
Toronto	0.965807	Minnesota	0.811810
Chicago	0.562178	Cleveland	0.960271
Detroit	0.253574	Kansas City	0.326836
Oakland	0.288287	Los Angeles	0.895267
Texas	0.500879	Seattle	0.839071
Houston	0.713682		

2. **Randomly Sampling PGA Golfers.** The U.S. Golf Association is considering a ban on long and belly putters. This has caused a great deal of controversy among both amateur golfers and members of the Professional Golf Association (PGA). Shown below are the names of the top 10 finishers in the recent PGA Tour McGladrey Classic golf tournament.

1. Tommy Gainey
2. David Toms
3. Jim Furyk
4. Brendon de Jonge
5. D. J. Trahan
6. Davis Love III
7. Chad Campbell
8. Greg Owens
9. Charles Howell III
10. Arjun Atwal

Select a simple random sample of 3 of these players to assess their opinions on the use of long and belly putters. LO 1

3. **Monthly Sales Data.** A simple random sample of 5 months of sales data provided the following information. LO 2, 3

Month:	1	2	3	4	5
Units Sold:	94	100	85	94	92

   a. Develop a point estimate of the population mean number of units sold per month.
   b. Develop a point estimate of the population standard deviation.

DATA*file* morningstar

4. **Morningstar Stock Data.** Morningstar publishes ratings data on 1,208 company stocks. A sample of 40 of these stocks is contained in the file *morningstar*. Use the Morningstar data set to answer the following questions. LO 3
   a. Develop a point estimate of the proportion of the stocks that receive Morningstar's highest rating of 5 Stars.
   b. Develop a point estimate of the proportion of the Morningstar stocks that are rated Above Average with respect to business risk.
   c. Develop a point estimate of the proportion of the Morningstar stocks that are rated 2 Stars or less.
5. **Internet Usage by Age Group.** One of the questions in the Pew Internet & American Life Project asked adults if they used the Internet at least occasionally. The results showed that 454 out of 478 adults aged 18–29 answered Yes; 741 out of 833

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

adults aged 30–49 answered Yes; and 1,058 out of 1,644 adults aged 50 and over answered Yes. LO 3

a. Develop a point estimate of the proportion of adults aged 18–29 who use the Internet.
b. Develop a point estimate of the proportion of adults aged 30–49 who use the Internet.
c. Develop a point estimate of the proportion of adults aged 50 and over who use the Internet.
d. Comment on any apparent relationship between age and Internet use.
e. Suppose your target population of interest is that of all adults (18 years of age and over). Develop an estimate of the proportion of that population who use the Internet.

DATA *file* eai

6. **Point Estimates for EAI Employees.** In this chapter we showed how a simple random sample of 30 EAI employees can be used to develop point estimates of the population mean annual salary, the population standard deviation for annual salary, and the population proportion having completed the management training program. LO 1, 2, 3, 4
   a. Use Excel to select a simple random sample of 50 EAI employees.
   b. Develop a point estimate of the mean annual salary.
   c. Develop a point estimate of the population standard deviation for annual salary.
   d. Develop a point estimate of the population proportion having completed the management training program.
7. **SAT Scores.** The College Board reported the following mean scores for the three parts of the SAT:
   Assume that the population standard deviation on each part of the test is $\sigma = 100$. LO 5

Critical Reading	502
Mathematics	515
Writing	494

   a. For a random sample of 30 test takers, what is the sampling distribution of $\bar{x}$ for scores on the Critical Reading part of the test?
   b. For a random sample of 60 test takers, what is the sampling distribution of $\bar{x}$ for scores on the Mathematics part of the test?
   c. For a random sample of 90 test takers, what is the sampling distribution of $\bar{x}$ for scores on the Writing part of the test?
8. **Federal Income Tax Returns.** *The Wall Street Journal* reports that 33% of taxpayers with adjusted gross incomes between $30,000 and $60,000 itemized deductions on their federal income tax return. The mean amount of deductions for this population of taxpayers was $16,642. Assume that the standard deviation is $\sigma = \$2{,}400$. LO 5
   a. What are the sampling distributions of $\bar{x}$ for itemized deductions for this population of taxpayers for each of the following sample sizes: 30, 50, 100, and 400?
   b. What is the advantage of a larger sample size when attempting to estimate the population mean?
9. **College Graduate-Level Wages.** ZipRecruiter periodically issues reports on wages of entry-level workers. They reported that entry-level wages for college graduates who took jobs in Deer Park, California were $23.69 per hour and for college graduates who took jobs in Stamford, Connecticut were $22.44 per hour in 2022. Assume that the standard deviation for graduates who took jobs in Deer Park, California is $2.30 and for graduates who took jobs in Stamford, Connecticut is $2.05. LO 5
   a. What is the sampling distribution of $\bar{x}$ for a random sample of 50 college graduates who took jobs in Deer Park, California?
   b. What is the sampling distribution of $\bar{x}$ for a random sample of 50 college graduates who took jobs in Stamford, Connecticut?
   c. In which of the preceding two cases, part (a) or part (b), is the standard error of $\bar{x}$ smaller? Why?
10. **State Rainfalls.** The state of California has a mean annual rainfall of 22 inches, whereas the state of New York has a mean annual rainfall of 42 inches. Assume that

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the standard deviation for both states is 4 inches. A sample of 30 years of rainfall for California and a sample of 45 years of rainfall for New York has been taken. LO 5

a. Show the sampling distribution of the sample mean annual rainfall for California.
b. Show the sampling distribution of the sample mean annual rainfall for New York.
c. In which of the preceding two cases, part (a) or part (b), is the standard error of $\overline{x}$ smaller? Why?

11. **Orders from First-Time Customers.** The president of Doerman Distributors, Inc. believes that 30% of the firm's orders come from first-time customers. A random sample of 100 orders will be used to estimate the proportion of first-time customers. Assume that the president is correct and $p = 0.30$. What is the sampling distribution of $\overline{p}$ for this study? LO 5

12. **Ages of Entrepreneurs.** *The Wall Street Journal* reported that the age at first startup for 55% of entrepreneurs was 29 years of age or less and the age at first startup for 45% of entrepreneurs was 30 years of age or more. LO 5
    a. Suppose a sample of 200 entrepreneurs will be taken to learn about the most important qualities of entrepreneurs. Show the sampling distribution of $\overline{p}$ where $\overline{p}$ is the sample proportion of entrepreneurs whose first startup was at 29 years of age or less.
    b. Suppose a sample of 200 entrepreneurs will be taken to learn about the most important qualities of entrepreneurs. Show the sampling distribution of $\overline{p}$ where $\overline{p}$ is now the sample proportion of entrepreneurs whose first startup was at 30 years of age or more.
    c. Are the standard errors of the sampling distributions different in parts (a) and (b)?

13. **Food Waste.** People end up tossing 12% of what they buy at the grocery store. Assume this is the true population proportion and that you plan to take a sample survey of 540 grocery shoppers to further investigate their behavior. Show the sampling distribution of $\overline{p}$, the proportion of groceries thrown out by your sample respondents. LO 5

14. **Unnecessary Medical Care.** Forty-two percent of primary care doctors think their patients receive unnecessary medical care. LO 5
    a. Suppose a sample of 300 primary care doctors was taken. Show the distribution of the sample proportion of doctors who think their patients receive unnecessary medical care.
    b. Suppose a sample of 500 primary care doctors was taken. Show the distribution of the sample proportion of doctors who think their patients receive unnecessary medical care.
    c. Suppose a sample of 1,000 primary care doctors was taken. Show the distribution of the sample proportion of doctors who think their patients receive unnecessary medical care.
    d. In which of the preceding three cases, part (a) or part (b) or part (c), is the standard error of $\overline{p}$ smallest? Why?

15. **Quality Ratings of Airports.** The International Air Transport Association surveys business travelers to develop quality ratings for transatlantic gateway airports. The maximum possible rating is 10. Suppose a simple random sample of 50 business travelers is selected and each traveler is asked to provide a rating for the Miami International Airport. The ratings obtained from the sample of 50 business travelers follow.

DATA*file* miami

6	4	6	8	7	7	6	3	3	8	10	4	8
7	8	7	5	9	5	8	4	3	8	5	5	4
4	4	8	4	5	6	2	5	9	9	8	4	8
9	9	5	9	7	8	3	10	8	9	6		

Develop a 95% confidence interval estimate of the population mean rating for Miami. LO 3

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

16. **Years to Bond Maturity.** A sample containing years to maturity and yield for 40 corporate bonds is contained in the file *corporatebonds*. LO 2.

DATA *file* corporatebonds

a. What is the sample mean years to maturity for corporate bonds and what is the sample standard deviation?
b. Develop a 95% confidence interval for the population mean years to maturity.
c. What is the sample mean yield on corporate bonds and what is the sample standard deviation?
d. Develop a 95% confidence interval for the population mean yield on corporate bonds.

17. **Telemedicine.** Health insurers are beginning to offer telemedicine services online that replace the common office visit. WellPoint provides a video service that allows subscribers to connect with a physician online and receive prescribed treatments. Wellpoint claims that users of its LiveHealth Online service saved a significant amount of money on a typical visit. The data shown below ($), for a sample of 20 online doctor visits, are consistent with the savings per visit reported by Wellpoint.

DATA *file* telehealth

92	34	40
105	83	55
56	49	40
76	48	96
93	74	73
78	93	100
53	82	

Assuming that the population is roughly symmetric, construct a 95% confidence interval for the mean savings for a televisit to the doctor as opposed to an office visit. LO 2

18. **Automobile Insurance Premiums.** The average annual premium for automobile insurance in the United States is $1,503. The following annual premiums ($) are representative of the web site's findings for the state of Michigan.

DATA *file* autoinsurance

1,905	3,112	2,312
2,725	2,545	2,981
2,677	2,525	2,627
2,600	2,370	2,857
2,962	2,545	2,675
2,184	2,529	2,115
2,332	2,442	

Assume the population is approximately normal. LO 2

a. Provide a point estimate of the mean annual automobile insurance premium in Michigan.
b. Develop a 95% confidence interval for the mean annual automobile insurance premium in Michigan.
c. Does the 95% confidence interval for the annual automobile insurance premium in Michigan include the national average for the United States? What is your interpretation of the relationship between auto insurance premiums in Michigan and the national average?

19. **Will Our Children Be Better Off?** One of the questions on a survey of 1,000 adults asked if today's children will be better off than their parents. Representative data are shown in the file *childoutlook*. A response of Yes indicates that the adult surveyed did think today's children will be better off than their parents. A response of No indicates that the adult surveyed did not think today's children will be better off than their parents. A response of Not Sure was given by 23% of the adults surveyed. LO 3

DATA *file* childoutlook

a. What is the point estimate of the proportion of the population of adults who do think that today's children will be better off than their parents?
b. At 95% confidence, what is the margin of error?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. What is the 95% confidence interval for the proportion of adults who do think that today's children will be better off than their parents?
d. What is the 95% confidence interval for the proportion of adults who do not think that today's children will be better off than their parents?
e. Which of the confidence intervals in parts (c) and (d) has the smaller margin of error? Why?

20. **Companies Exceeding Profit Estimates.** According to Thomson Financial, last year the majority of companies reporting profits had beaten estimates. A sample of 162 companies showed that 104 beat estimates, 29 matched estimates, and 29 fell short. LO 3
a. What is the point estimate of the proportion that fell short of estimates?
b. Determine the margin of error and provide a 95% confidence interval for the proportion that beat estimates.
c. How large a sample is needed if the desired margin of error is 0.05?

21. **Internet Usage.** The Pew Research Center Internet Project conducted a survey of 857 Internet users. This survey provided a variety of statistics on them. LO 3
a. The sample survey showed that 90% of respondents said the Internet has been a good thing for them personally. Develop a 95% confidence interval for the proportion of respondents who say the Internet has been a good thing for them personally.
b. The sample survey showed that 67% of Internet users said the Internet has generally strengthened their relationship with family and friends. Develop a 95% confidence interval for the proportion of respondents who say the Internet has strengthened their relationship with family and friends.
c. Fifty-six percent of Internet users have seen an online group come together to help a person or community solve a problem, whereas only 25% have left an online group because of unpleasant interaction. Develop a 95% confidence interval for the proportion of Internet users who say online groups have helped solve a problem.
d. Compare the margin of error for the interval estimates in parts (a), (b), and (c). How is the margin of error related to the sample proportion?

22. **Employee Contributions to Health-Care Coverage.** For many years businesses have struggled with the rising cost of healthcare. But recently, the increases have slowed due to less inflation in health-care prices and employees paying for a larger portion of health-care benefits. A recent Mercer survey showed that 52% of U.S. employers were likely to require higher employee contributions for health-care coverage. Suppose the survey was based on a sample of 800 companies. Compute the margin of error and a 95% confidence interval for the proportion of companies likely to require higher employee contributions for health-care coverage. LO 3

23. **Hotel Guest Bills.** The manager of the Danvers-Hilton Resort Hotel stated that the mean guest bill for a weekend is $600 or less. A member of the hotel's accounting staff noticed that the total charges for guest bills have been increasing in recent months. The accountant will use a sample of future weekend guest bills to test the manager's claim. LO 6
a. Which form of the hypotheses should be used to test the manager's claim? Explain.

$$H_0: \mu \geq 600 \quad H_0: \mu \leq 600 \quad H_0: \mu = 600$$
$$H_a: \mu < 600 \quad H_a: \mu > 600 \quad H_a: \mu \neq 600$$

b. What conclusion is appropriate when H_0 cannot be rejected?
c. What conclusion is appropriate when H_0 can be rejected?

24. **Bonus Plans and Automobile Sales.** The manager of an automobile dealership is considering a new bonus plan designed to increase sales volume. Currently, the mean sales volume is 14 automobiles per month. The manager wants to conduct a research study to see whether the new bonus plan increases sales volume. To collect data on the plan, a sample of sales personnel will be allowed to sell under the new bonus plan for a one-month period. LO 6

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Develop the null and alternative hypotheses most appropriate for this situation.
b. Comment on the conclusion when H_0 cannot be rejected.
c. Comment on the conclusion when H_0 can be rejected.

25. **Filling Detergent Cartons.** A production line operation is designed to fill cartons with laundry detergent to a mean weight of 32 ounces. A sample of cartons is periodically selected and weighed to determine whether underfilling or overfilling is occurring. If the sample data lead to a conclusion of underfilling or overfilling, the production line will be shut down and adjusted to obtain proper filling. **LO 6**
    a. Formulate the null and alternative hypotheses that will help in deciding whether to shut down and adjust the production line.
    b. Comment on the conclusion and the decision when H_0 cannot be rejected.
    c. Comment on the conclusion and the decision when H_0 can be rejected.

26. **Process Improvement.** Because of high production-changeover time and costs, a director of manufacturing must convince management that a proposed manufacturing method reduces costs before the new method can be implemented. The current production method operates with a mean cost of $220 per hour. A research study will measure the cost of the new method over a sample production period. **LO 6**
    a. Develop the null and alternative hypotheses most appropriate for this study.
    b. Comment on the conclusion when H_0 cannot be rejected.
    c. Comment on the conclusion when H_0 can be rejected.

27. **Home Electricity Usage.** Duke Energy reported that the cost of electricity for an efficient home in a particular neighborhood of Cincinnati, Ohio, was $104 per month. A researcher believes that the cost of electricity for a comparable neighborhood in Chicago, Illinois, is higher. A sample of homes in this Chicago neighborhood will be taken and the sample mean monthly cost of electricity will be used to test the following null and alternative hypotheses. **LO 6, 9**

$$H_0: \mu \leq 104$$
$$H_a: \mu > 104$$

    a. Assume that the sample data lead to rejection of the null hypothesis. What would be your conclusion about the cost of electricity in the Chicago neighborhood?
    b. What is the Type I error in this situation? What are the consequences of making this error?
    c. What is the Type II error in this situation? What are the consequences of making this error?

28. **Orange Juice Labels.** The label on a 3-quart container of orange juice states that the orange juice contains an average of 1 gram of fat or less. Answer the following questions for a hypothesis test that could be used to test the claim on the label. **LO 6, 9**
    a. Develop the appropriate null and alternative hypotheses.
    b. What is the Type I error in this situation? What are the consequences of making this error?
    c. What is the Type II error in this situation? What are the consequences of making this error?

29. **Carpet Salesperson Salaries.** Carpetland salespersons average $8,000 per week in sales. Steve Contois, the firm's vice president, proposes a compensation plan with new selling incentives. Steve hopes that the results of a trial selling period will lead to the conclusion that the compensation plan increases the average sales per salesperson. **LO 6, 9**
    a. Develop the appropriate null and alternative hypotheses.
    b. What is the Type I error in this situation? What are the consequences of making this error?
    c. What is the Type II error in this situation? What are the consequences of making this error?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

30. **Production Operating Costs.** Suppose a new production method will be implemented if a hypothesis test supports the conclusion that the new method reduces the mean operating cost per hour. **LO 6, 9**
    a. State the appropriate null and alternative hypotheses if the mean cost for the current production method is $220 per hour.
    b. What is the Type I error in this situation? What are the consequences of making this error?
    c. What is the Type II error in this situation? What are the consequences of making this error?

31. **Meal Costs.** Which is cheaper: eating out or dining in? The mean cost of a flank steak, broccoli, and rice bought at the grocery store is $23.04. A sample of 100 neighborhood restaurants showed a mean price of $22.75 and a standard deviation of $2 for a comparable restaurant meal. **LO 6, 7**
    a. Develop appropriate hypotheses for a test to determine whether the sample data support the conclusion that the mean cost of a restaurant meal is less than fixing a comparable meal at home.
    b. Using the sample from the 100 restaurants, what is the p value?
    c. At $\alpha = 0.05$, what is your conclusion?

32. **CEO Tenure.** A shareholders' group, in lodging a protest, claimed that the mean tenure for a chief executive officer (CEO) was at least nine years. A survey of companies reported in *The Wall Street Journal* found a sample mean tenure of $\bar{x} = 7.27$ years for CEOs with a standard deviation of $s = 6.38$ years. **LO 6, 7**
    a. Formulate hypotheses that can be used to challenge the validity of the claim made by the shareholders' group.
    b. Assume that 85 companies were included in the sample. What is the p value for your hypothesis test?
    c. At $\alpha = 0.01$, what is your conclusion?

33. **School Administrator Salaries.** Salary.com reports that national mean annual salary for a school administrator is $80,935 a year. A school official took a sample of 25 school administrators in the state of Ohio to learn about salaries in that state to see if they differed from the national average. **LO 6, 7**

DATA*file* administrator

    a. Formulate hypotheses that can be used to determine whether the population mean annual administrator salary in Ohio differs from the national mean of $90,000.
    b. The sample data for 25 Ohio administrators is contained in the file *administrator*. What is the p value for your hypothesis test in part (a)?
    c. At $\alpha = 0.05$, can your null hypothesis be rejected? What is your conclusion?

34. **Time in Child Care.** The time married men with children spend on child care averages 6.4 hours per week. You belong to a professional group on family practices that would like to do its own study to determine if the time married men in your area spend on child care per week differs from the reported mean of 6.4 hours per week. A sample of 40 married couples will be used with the data collected showing the hours per week the husband spends on child care. The sample data are contained in the file *childcare*. **LO 6, 7**

DATA*file* childcare

    a. What are the hypotheses if your group would like to determine if the population mean number of hours married men are spending on child care differs from the mean reported by *Time* in your area?
    b. What is the sample mean and the p value?
    c. Select your own level of significance. What is your conclusion?

35. **Per Capita Sales.** The Coca-Cola Company reported that the mean per capita annual sales of its beverages in the United States was 423 eight-ounce servings. Suppose you are curious whether the consumption of Coca-Cola beverages is higher in Atlanta, Georgia, the location of Coca-Cola's corporate headquarters. A sample of 36 individuals from the Atlanta area showed a sample mean annual consumption of 460.4 eight-ounce servings with a standard deviation of $s = 101.9$ ounces. Using $\alpha = 0.05$, do the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

sample results support the conclusion that mean annual consumption of Coca-Cola beverage products is higher in Atlanta? LO 6, 7

36. **Used Car Prices.** According to Autoweek, the mean price for used cars is $33,341. A manager of a Kansas City used car dealership reviewed a sample of 50 recent used car sales at the dealership in an attempt to determine whether the population mean price for used cars at this particular dealership differed from the national mean. The prices for the sample of 50 cars are shown in the file *usedcars*. LO 6, 7.

DATA *file* usedcars

   a. Formulate the hypotheses that can be used to determine whether a difference exists in the mean price for used cars at the dealership.
   b. What is the p value?
   c. At $\alpha = 0.05$, what is your conclusion?

37. **Population Mobility.** What percentage of the population live in their state of birth? According to the U.S. Census Bureau's American Community Survey, the figure ranges from 25% in Nevada to 78.7% in Louisiana. The average percentage across all states and the District of Columbia is 57.7%. The data in the file *homestate* are consistent with the findings in the American Community Survey. The data are for a random sample of 120 Arkansas residents and for a random sample of 180 Virginia residents. LO 6, 8

DATA *file* homestate

   a. Formulate hypotheses that can be used to determine whether the percentage of stay-at-home residents in the two states differs from the overall average of 57.7%.
   b. Estimate the proportion of stay-at-home residents in Arkansas. Does this proportion differ significantly from the mean proportion for all states? Use $\alpha = 0.05$.
   c. Estimate the proportion of stay-at-home residents in Virginia. Does this proportion differ significantly from the mean proportion for all states? Use $\alpha = 0.05$.
   d. Would you expect the proportion of stay-at-home residents to be higher in Virginia than in Arkansas? Support your conclusion with the results obtained in parts (b) and (c).

38. **Holiday Gifts from Employers.** Last year, 46% of business owners gave a holiday gift to their employees. A survey of business owners indicated that 35% plan to provide a holiday gift to their employees. Suppose the survey results are based on a sample of 60 business owners. LO 6, 8
   a. How many business owners in the survey plan to provide a holiday gift to their employees?
   b. Suppose the business owners in the sample do as they plan. Compute the p value for a hypothesis test that can be used to determine if the proportion of business owners providing holiday gifts has decreased from last year.
   c. Using a 0.05 level of significance, would you conclude that the proportion of business owners providing gifts has decreased? What is the smallest level of significance for which you could draw such a conclusion?

39. **Family Stock Ownership.** Ten years ago 53% of American families owned stocks or stock funds. Sample data collected by the Investment Company Institute indicate that the percentage is now 46%. LO 6, 8
   a. Develop appropriate hypotheses such that rejection of H_0 will support the conclusion that a smaller proportion of American families own stocks or stock funds this year than 10 years ago.
   b. Assume that the Investment Company Institute sampled 300 American families to estimate that the percent owning stocks or stock funds is 46% this year. What is the p value for your hypothesis test?
   c. At $\alpha = 0.01$, what is your conclusion?

40. **Returned Merchandise.** According to the University of Nevada Center for Logistics Management, 6% of all merchandise sold in the United States gets returned. A Houston department store sampled 80 items sold in January and found that 12 of the items were returned. LO 3, 6, 8

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Construct a point estimate of the proportion of items returned for the population of sales transactions at the Houston store.
b. Construct a 95% confidence interval for the proportion of returns at the Houston store.
c. Is the proportion of returns at the Houston store significantly different from the returns for the nation as a whole? Provide statistical support for your answer.

41. **Coupon Usage.** Eagle Outfitters is a chain of stores specializing in outdoor apparel and camping gear. It is considering a promotion that involves mailing discount coupons to all its credit card customers. This promotion will be considered a success if more than 10% of those receiving the coupons use them. Before going national with the promotion, coupons were sent to a sample of 100 credit card customers. **LO 3, 6, 8**

DATA*file* eagle

a. Develop hypotheses that can be used to test whether the population proportion of those who will use the coupons is sufficient to go national.
b. The file *eagle* contains the sample data. Develop a point estimate of the population proportion.
c. Use $\alpha = 0.05$ to conduct your hypothesis test. Should Eagle go national with the promotion?

42. **Malpractice Suits.** One of the reasons health-care costs have been rising rapidly in recent years is the increasing cost of malpractice insurance for physicians. Also, fear of being sued causes doctors to run more precautionary tests (possibly unnecessary) just to make sure they are not guilty of missing something. These precautionary tests also add to health-care costs. Data in the file *lawsuit* are consistent with findings in a *Reader's Digest* article and can be used to estimate the proportion of physicians over the age of 55 who have been sued at least once. **LO 3, 6, 8**

DATA*file* lawsuit

a. Formulate hypotheses that can be used to see if these data can support a finding that more than half of physicians over the age of 55 have been sued at least once.
b. Use Excel and the file *lawsuit* to compute the sample proportion of physicians over the age of 55 who have been sued at least once. What is the *p* value for your hypothesis test?
c. At $\alpha = 0.01$, what is your conclusion?

43. **Value of Orders Placed.** The Port Authority sells a wide variety of cables and adapters for electronic equipment online. Last year the mean value of orders placed with the Port Authority was $47.28, and management wants to assess whether the mean value of orders placed to date this year is the same as last year. The values of a sample of 49,896 orders placed this year are collected and recorded in the file *portauthority*. **LO 6, 7**

DATA*file* portauthority

a. Formulate hypotheses that can be used to test whether the mean value of orders placed this year differs from the mean value of orders placed last year.
b. Use the data in the file *portauthority* to conduct your hypothesis test. What is the *p* value for your hypothesis test? At $\alpha = 0.01$, what is your conclusion?

44. **Customer Gender.** The Port Authority also wants to determine if the gender profile of its customers has changed since last year, when 59.4% of its orders placed were placed by males. The genders for a sample of 49,896 orders placed this year are collected and recorded in the file *portauthority*. **LO 6, 8**.

DATA*file* portauthority

a. Formulate hypotheses that can be used to test whether the proportion of orders placed by male customers this year differs from the proportion of orders placed by male customers placed last year.
b. Use the data in the file *portauthority* to conduct your hypothesis test. What is the *p* value for your hypothesis test? At $\alpha = 0.05$, what is your conclusion?

45. **Erroneous Federal Tax Returns.** Suppose a sample of 10,001 erroneous Federal income tax returns from last year has been taken and is provided in the file *fedtaxerrors*. A positive value indicates the taxpayer underpaid and a negative value indicates that the taxpayer overpaid. **LO 2, 6, 7**

DATA*file* fedtaxerrors

a. What is the sample mean error made on erroneous Federal income tax returns last year?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Using 95% confidence, what is the margin of error?
c. Using the results from parts (a) and (b), develop the 95% confidence interval estimate of the mean error made on erroneous Federal income tax returns last year.

46. **Federal Employee Sick Days.** According to the Congressional Research Service, 2,191,011 people are employed by the federal government in the United States. Suppose that a random sample of 3,500 of these federal employees was selected and the number of sick hours each of these employees took last year was collected from an electronic personnel database. The data collected in this survey are provided in the file *fedsickhours.* **LO 2**



a. What is the sample mean number of sick hours taken by federal employees last year?
b. Using 99% confidence, what is the margin of error?
c. Using the results from parts (a) and (b), develop the 99% confidence interval estimate of the mean number of sick hours taken by federal employees last year.
d. If the mean sick hours federal employees took two years ago was 62.2, what would the confidence interval in part (c) lead you to conclude about last year?

47. **Web Browser Preference.** Internet users were recently asked online to rate their satisfaction with the web browser they use most frequently. Of 102,519 respondents, 65,120 indicated they were very satisfied with the web browser they use most frequently. **LO 3**
a. What is the sample proportion of Internet users who are very satisfied with the web browser they use most frequently?
b. Using 95% confidence, what is the margin of error?
c. Using the results from parts (a) and (b), develop the 95% confidence interval estimate of the proportion of Internet users who are very satisfied with the web browser they use most frequently.

48. **Speeding Drivers.** ABC News reports that 58% of U.S. drivers admit to speeding. Suppose that a new satellite technology can instantly measure the speed of any vehicle on a U.S. road and determine whether the vehicle is speeding, and this satellite technology was used to take a sample of 20,000 vehicles at 6:00 p.m. EST on a recent Tuesday afternoon. Of these 20,000 vehicles, 9,252 were speeding. **LO 3**
a. What is the sample proportion of vehicles on U.S. roads that speed?
b. Using 99% confidence, what is the margin of error?
c. Using the results from parts (a) and (b), develop the 99% confidence interval estimate of the proportion of vehicles on U.S. roads that speed.
d. What does the confidence interval in part (c) lead you to conclude about the ABC News report?

49. **Government Use of E-mail.** The Federal Government wants to determine if the mean number of business e-mails sent and received per business day by its employees differs from the mean number of e-mails sent and received per day by corporate employees, which is 101.5. Suppose the department electronically collects information on the number of business e-mails sent and received on a randomly selected business day over the past year from each of 10,163 randomly selected Federal employees. The results are provided in the file *fedemail*. Test the Federal Government's hypothesis at $\alpha = 0.01$. Discuss the practical significance of the results. **LO 2, 6, 7**



50. **CEOs and Social Networks.** CEOs who belong to a popular business-oriented social networking service have an average of 930 connections. Do other members have fewer connections than CEOs? The number of connections for a random sample of 7,515 members who are not CEOs is provided in the file *socialnetwork*. Using this sample, test the hypothesis that other members have fewer connections than CEOs at $\alpha = 0.01$. Discuss the practical significance of the results. **LO 2, 6, 7**



51. **French Fry Purchases.** The American Potato Growers Association (APGA) would like to test the claim that the proportion of fast-food orders this year that include French fries exceeds the proportion of fast-food orders that included French fries last year. Suppose that a random sample of 49,581 electronic receipts for fast-food orders placed this year shows that 31,038 included French fries. Assuming that the proportion of

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

fast-food orders that included French fries last year is 0.62, use this information to test APGA's claim at $\alpha = 0.05$. Discuss the practical significance of the results. **LO 3, 6, 8**

52. **GPS Usage in Canada.** According to CNN, 55% of all U.S. smartphone users have used their GPS capability to get directions. Suppose a major provider of wireless telephone service in Canada wants to know how GPS usage by its customers compares with U.S. smartphone users. The company collects usage records for this year for a random sample of 547,192 of its Canadian customers and determines that 302,050 of these customers have used their telephone's GPS capability this year. Use this data to test whether Canadian smartphone users' GPS usage differs from U.S. smartphone users' GPS usage at $\alpha = 0.01$. Discuss the practical significance of the results. **LO 6, 8**

53. **Election Poll.** A well-respected polling agency has conducted a poll for an upcoming Presidential election. The polling agency has taken measures so that its random sample consists of 50,000 people and is representative of the voting population. The file *pedro* contains survey data for 50,000 respondents in both a pre-election survey and a post-election poll. **LO 3**

DATA*file*
pedro

a. Based on the data in the "Support Pedro in Pre-Election Poll" column, compute the 99% confidence interval on the population proportion of voters who support Pedro Ringer in the upcoming election. If Pedro needs at least 50% of the vote to win in the two-party election, should Pedro be optimistic about winning the election?
b. Now suppose the election occurs and Pedro wins 55% of the vote. Explain how this result could occur given the sample information in part (a).
c. In an attempt to explain the election results (Pedro winning 55% of the vote), the polling agency has followed up with each of the respondents in their pre-election survey. The data in the "Voted for Pedro?" column corresponds to whether or not the respondent actually voted for Pedro in the election. Compute the 99% confidence interval on the population proportion of voters who voted for Pedro Ringer. Is this result consistent with the election results?
d. Use a PivotTable to determine the percentage of survey respondents who voted for Pedro that did not admit to supporting Pedro in a pre-election poll. Use this result to explain the discrepancy between the pre-election poll and the actual election results. What type of error is occurring here?

Case Problem 1: Young Professional Magazine

Young Professional magazine was developed for a target audience of recent college graduates who are in their first 10 years in a business/professional career. In its two years of publication, the magazine has been fairly successful. Now the publisher is interested in expanding the magazine's advertising base. Potential advertisers continually ask about the demographics and interests of subscribers to *Young Professional*. To collect this information, the magazine commissioned a survey to develop a profile of its subscribers. The survey results will be used to help the magazine choose articles of interest and provide advertisers with a profile of subscribers. As a new employee of the magazine, you have been asked to help analyze the survey results, a portion of which are shown in the following table.

Age	Sex	Real Estate Purchases	Value of Investments ($)	Number of Transactions	Broadband Access	Household Income ($)	Children
38	Female	No	12,200	4	Yes	75,200	Yes
30	Male	No	12,400	4	Yes	70,300	Yes
41	Female	No	26,800	5	Yes	48,200	No
28	Female	Yes	19,600	6	No	95,300	No
31	Female	Yes	15,100	5	No	73,300	Yes
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file* professional

Some of the survey questions are as follows:

1. What is your age?
2. Are you: Male _____ Female _____
3. Do you plan to make any real estate purchases in the next two years? Yes_____ No_____
4. What is the approximate total value of financial investments, exclusive of your home, owned by you or members of your household?
5. How many stock/bond/mutual fund transactions have you made in the past year?
6. Do you have broadband access to the Internet at home? Yes_____ No_____
7. Please indicate your total household income last year.
8. Do you have children? Yes_____ No_____

The file *professional* contains the responses to these questions. The table shows the portion of the file pertaining to the first five survey respondents. **LO 2, 3**

Managerial Report

Prepare a managerial report summarizing the results of the survey. In addition to statistical summaries, discuss how the magazine might use these results to attract advertisers. You might also comment on how the survey results could be used by the magazine's editors to identify topics that would be of interest to readers. Your report should address the following issues, but do not limit your analysis to just these areas.

1. Develop appropriate descriptive statistics to summarize the data.
2. Develop 95% confidence intervals for the mean age and household income of subscribers.
3. Develop 95% confidence intervals for the proportion of subscribers who have broadband access at home and the proportion of subscribers who have children.
4. Would *Young Professional* be a good advertising outlet for online brokers? Justify your conclusion with statistical data.
5. Would this magazine be a good place to advertise for companies selling educational software and computer games for young children?
6. Comment on the types of articles you believe would be of interest to readers of *Young Professional*.

Case Problem 2: Quality Associates, Inc.

Quality Associates, Inc., a consulting firm, advises its clients about sampling and statistical procedures that can be used to control their manufacturing processes. In one particular application, a client gave Quality Associates a sample of 800 observations taken while that client's process was operating satisfactorily. The sample standard deviation for these data was 0.21; hence, with so much data, the population standard deviation was assumed to be 0.21. Quality Associates then suggested that random samples of size 30 be taken periodically to monitor the process on an ongoing basis. By analyzing the new samples, the client could quickly learn whether the process was operating satisfactorily. When the process was not operating satisfactorily, corrective action could be taken to eliminate the problem. The design specification indicated that the mean for the process should be 12. The hypothesis test suggested by Quality Associates is as follows:

$$H_0: \mu = 12$$
$$H_a: \mu \neq 12$$

Corrective action will be taken any time H_0 is rejected.

The samples listed in the following table were collected at hourly intervals during the first day of operation of the new statistical process control procedure. These data are available in the file *quality* **LO 2, 4, 7**.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
quality

Sample 1	Sample 2	Sample 3	Sample 4
11.55	11.62	11.91	12.02
11.62	11.69	11.36	12.02
11.52	11.59	11.75	12.05
11.75	11.82	11.95	12.18
11.90	11.97	12.14	12.11
11.64	11.71	11.72	12.07
11.80	11.87	11.61	12.05
12.03	12.10	11.85	11.64
11.94	12.01	12.16	12.39
11.92	11.99	11.91	11.65
12.13	12.20	12.12	12.11
12.09	12.16	11.61	11.90
11.93	12.00	12.21	12.22
12.21	12.28	11.56	11.88
12.32	12.39	11.95	12.03
11.93	12.00	12.01	12.35
11.85	11.92	12.06	12.09
11.76	11.83	11.76	11.77
12.16	12.23	11.82	12.20
11.77	11.84	12.12	11.79
12.00	12.07	11.60	12.30
12.04	12.11	11.95	12.27
11.98	12.05	11.96	12.29
12.30	12.37	12.22	12.47
12.18	12.25	11.75	12.03
11.97	12.04	11.96	12.17
12.17	12.24	11.95	11.94
11.85	11.92	11.89	11.97
12.30	12.37	11.88	12.23
12.15	12.22	11.93	12.25

Managerial Report

1. Conduct a hypothesis test for each sample at the 0.01 level of significance and determine what action, if any, should be taken. Provide the test statistic and p value for each test.
2. Compute the standard deviation for each of the four samples. Does the conjecture of 0.21 for the population standard deviation appear reasonable?
3. Compute limits for the sample mean $\bar{x}$ around $\mu = 12$ such that, as long as a new sample mean is within those limits, the process will be considered to be operating satisfactorily. If $\bar{x}$ exceeds the upper limit or if $\bar{x}$ is below the lower limit, corrective action will be taken. These limits are referred to as upper and lower control limits for quality-control purposes.
4. Discuss the implications of changing the level of significance to a larger value. What mistake or error could increase if the level of significance is increased?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 8

Linear Regression

Contents

Analytics in Action: Walmart.com

8.1 Simple Linear Regression Model
- Estimated Simple Linear Regression Equation

8.2 Least Squares Method
- Least Squares Estimates of the Simple Linear Regression Parameters
- Using Excel's Chart Tools to Compute the Estimated Simple Linear Regression Equation

8.3 Assessing the Fit of the Simple Linear Regression Model
- The Sums of Squares
- The Coefficient of Determination
- Using Excel's Chart Tools to Compute the Coefficient of Determination

8.4 The Multiple Linear Regression Model
- Estimated Multiple Linear Regression Equation
- Least Squares Method and Multiple Linear Regression
- Butler Trucking Company and Multiple Linear Regression
- Using Excel's Regression Tool to Develop the Estimated Multiple Linear Regression Equation

8.5 Inference and Linear Regression
- Conditions Necessary for Valid Inference in the Least Squares Linear Regression Model
- Testing Individual Linear Regression Parameters
- Addressing Nonsignificant Independent Variables
- Multicollinearity

8.6 Categorical Independent Variables
- Butler Trucking Company and Rush Hour
- Interpreting the Parameters
- More Complex Categorical Variables

8.7 Modeling Nonlinear Relationships
- Quadratic Regression Models
- Piecewise Linear Regression Models
- Interaction Between Independent Variables

8.8 Model Fitting
- Variable Selection Procedures
- Overfitting

8.9 Big Data and Linear Regression
- Inference and Very Large Samples
- Model Selection

8.10 Prediction with Linear Regression

Summary 450
Glossary 450
Problems 452

Available in the Cengage eBook:
- Appendix: Simple Linear Regression with R
- Appendix: Multiple Linear Regression with R
- Appendix: Linear Regression Variable Selection Procedures with R

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Construct an estimated simple linear regression model that estimates how a dependent variable is related to an independent variable.

LO 2 Construct an estimated multiple linear regression model that estimates how a dependent variable is related to multiple independent variables.

LO 3 Compute and interpret the estimated coefficient of determination for a linear regression model.

LO 4 Assess whether the conditions necessary for valid inference in a least squares linear regression model are satisfied and test hypotheses about the parameters.

LO 5 Test hypotheses about the parameters of a linear regression model and interpret the results of these hypotheses tests.

LO 6 Compute and interpret confidence intervals for the parameters of a linear regression model.

LO 7 Use dummy variables to incorporate categorical independent variables in a linear regression model and interpret the associated estimated regression parameters.

LO 8 Use a quadratic regression model, a piecewise linear regression model, and interaction between independent variables to account for curvilinear relationships between independent variables and the dependent variable in a regression model and interpret the estimated parameters.

LO 9 Use an estimated linear regression model to predict the value of the dependent variable given values of the independent variables.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

Walmart.com*

Bentonville, Arkansas

With more than 245 million customers per week visiting its 11,000 stores across the globe, Walmart is the world's largest retailer. In 2000, in response to increasing online shopping, Walmart launched its Internet site *Walmart.com*. To serve its online customers, Walmart created a network of distribution centers in the United States. One of these distribution centers, located in Carrollton, Georgia, was selected as the site for a study on how Walmart might better manage its packaging for online orders.

In its peak shipping season (November to December), the Walmart distribution center in Carrollton ships over 100,000 packages per day. The cost of fulfilling an order includes the material cost (the cost of the carton and packing material—paper that fills the empty space when the product is smaller than the volume of the box), labor for handling, and the shipping cost. The study conducted in Carrollton, called the Carton-mix Optimization Study, had as its objective to determine the number and size of cartons to have on hand for shipping Walmart's products to its online customers that would minimize raw material, labor and shipping costs.

A number of constraints limited the possibilities for the variety of cartons to have on hand. For example, management provided a minimum and maximum carton size. In addition, for boxes automatically constructed (as opposed to manually built), a line would be limited to one-size carton.

As with any study of this type, data had to be collected to build a cost model that could then be used to find the optimal carton mix. The material costs for existing cartons were known, but what about the cost of a carton in a size that is not currently used? Based on price quotes from its carton suppliers for a variety of sizes, a simple linear regression model was used to estimate the cost of any-size carton based on its volume. The following model was used to estimate the material cost of a carton in dollars per carton (y) based on the volume of the carton measured in cubic inches per carton (x):

$$y = -0.11 + 0.0014x$$

For example, for a carton with volume of 2,800 cubic inches, we have $-0.11 + 0.0014(2800) = 3.81$. Therefore, the estimated material cost for a carton of 2,800 cubic inches in volume is \$3.81.

The simple linear regression model was imbedded into an optimization algorithm in Microsoft Excel to provide Walmart managers a recommendation on the optimal mix of carton sizes to carry at its Carrollton distribution center. The optimization model based on this simple linear regression provided documented savings of \$600,000 in its first year of implementation. The model was later deployed across Walmart's other distribution centers with an estimated annual savings of \$2 million.

In this chapter, we will study how to estimate a simple linear regression model, that is, a linear model in which a single variable is used to estimate the value of another variable.

*Based on S. Ahire, M. Malhotra, and J. Jensen, "Carton-Mix Optimization for Walmart.com Distribution Centers," *Interfaces*, Vol. 45, No. 4, July–August 2015, pp. 341–357.

Managerial decisions are often based on the relationship between two or more variables. For example, after considering the relationship between advertising expenditures and sales, a marketing manager might attempt to predict sales for a given level of advertising expenditures. In another case, a public utility might use the relationship between the daily high temperature and the demand for electricity to predict electricity usage on the basis of next month's anticipated daily high temperatures. The variable being predicted is called the **dependent variable** (also referred to as the *response* or *target* variable) and the variables being used to predict the value of the dependent variable are called the **independent variables** (also referred to as *predictor variables* or *features*). For example, in analyzing the effect of advertising expenditures on sales, a marketing manager's desire to predict sales would suggest making sales the dependent variable. Advertising expenditure would be the independent variable used to help predict sales.

In Chapter 10, we discuss other methods for addressing the task of predicting a quantitative variable besides linear regression.

In Chapter 11, we discuss methods for addressing the task of predicting a variable when it is categorical instead of quantitative.

When the variable to be predicted is a quantitative, the prediction task is called **regression**. In this chapter, we begin by considering **simple linear regression**; in this term, "simple" means we are looking at the relationship between one dependent variable (denoted by y) and one independent variable (denoted by x), and "linear" means this relationship is approximated by a straight line. We then extend this concept to higher dimensions by introducing **multiple linear regression** to model the relationship between a dependent variable (y) and two or more independent variables ($x_1, x_2, \ldots, x_q$).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8.1 Simple Linear Regression Model

Butler Trucking Company is an independent trucking company in Southern California. A major portion of Butler's business involves deliveries throughout its local area. To develop better work schedules, the managers want to estimate the total daily travel times for their drivers. The managers believe that the total daily travel times (denoted by y) are closely related to the number of miles traveled in making the daily deliveries (denoted by x). Using linear regression, we can develop an equation showing how the dependent variable y is related to the independent variable x.

In the Butler Trucking Company example, a simple linear regression model hypothesizes that the travel time of a driving assignment (y) is linearly related to the number of miles traveled (x) as follows:

Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon \quad \textbf{(8.1)}$$

We introduce the concept of a random variable in Chapter 7.

In equation (8.1), β_0 and β_1 are **population parameters**, or measurable factors that define characteristics of a population, process, or system. These parameters describe the y-intercept and slope of the line relating y and x. The error term ε (Greek letter epsilon) accounts for the variability in y that cannot be explained by the linear relationship between x and y. The simple linear regression model assumes that the error term is a normally distributed random variable with a mean of zero and constant variance for all observations.

Estimated Simple Linear Regression Equation

In practice, the values of the population parameters β_0 and β_1 are not known and must be estimated using sample data. Sample statistics (denoted b_0 and b_1) are computed as estimates of the population parameters β_0 and β_1. Substituting the values of the sample statistics b_0 and b_1 for β_0 and β_1 in equation (8.1) and dropping the error term (because its expected value is zero), we obtain the **estimated linear regression** for simple linear regression:

Estimated Simple Linear Regression Equation

$$\hat{y} = b_0 + b_1 x \quad \textbf{(8.2)}$$

Figure 8.1 provides a summary of the estimation process for simple linear regression. Using equation (8.2), $\hat{y}$ provides an estimate for the mean value of y corresponding to a given value of x.

The graph of the estimated simple linear regression equation is called the *estimated regression line*; b_0 is the estimated y-intercept, and b_1 is the estimated slope. In the next section, we show how the least squares method can be used to compute the values of b_0 and b_1 in the estimated linear regression equation.

Examples of possible regression lines are shown in Figure 8.2. The regression line in Panel A shows that the estimated mean value of y is related positively to x, with larger values of $\hat{y}$ associated with larger values of x. In Panel B, the estimated mean value of y is related negatively to x, with smaller values of $\hat{y}$ associated with larger values of x. In Panel C, the estimated mean value of y is not related to x; that is, $\hat{y}$ is the same for every value of x.

In general, $\hat{y}$ is the **point estimator** of $E(y|x)$, the mean value of y for a given value of x. Thus, to estimate the mean or expected value of travel time for a driving assignment of 75 miles, Butler Trucking would substitute the value of 75 for x in equation (8.2). In some cases, however, Butler Trucking may be more interested in predicting travel time for an upcoming driving assignment of a particular length. For example, suppose Butler Trucking would like to predict travel time for a new 75-mile driving assignment the company is considering. It turns out that to predict travel time for a new 75-mile driving assignment, Butler Trucking would also substitute the value of 75 for x in equation (8.2). The value of $\hat{y}$ provides both a point estimate of $E(y|x)$ for a given value of x and a prediction of an individual value of y for a given value of x. In most cases, we will refer to $\hat{y}$ simply as the predicted value of y.

A point estimator is a single value used as an estimate of the corresponding population parameter.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The estimation of β_0 and β_1 is a statistical process much like the estimation of the population mean, μ, discussed in Chapter 7. β_0 and β_1 are the unknown parameters of interest, and b_0 and b_1 are the sample statistics used to estimate the parameters.

Figure 8.1 The Estimation Process in Simple Linear Regression

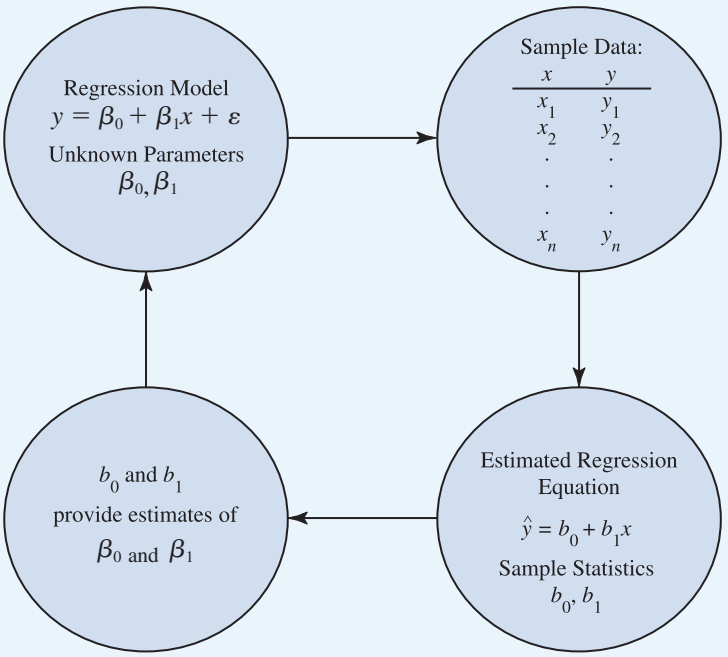
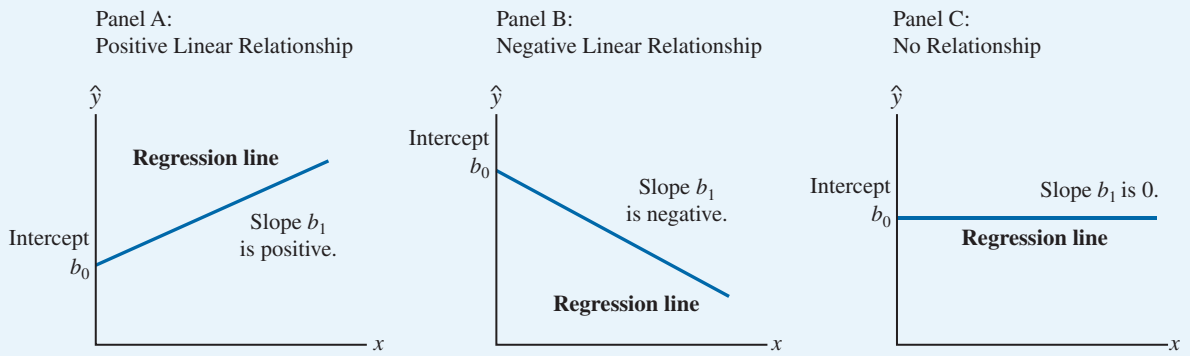


Figure 8.2 Possible Regression Lines in Simple Linear Regression



Notes + Comments

1. The statistical methods used in studying the relationship between two variables were first employed by Sir Francis Galton (1822–1911). Galton found that the heights of the sons of unusually tall or unusually short fathers tend to move, or "regress," toward the average height of the male population. Karl Pearson 1857–1936), a disciple of Galton, later confirmed this finding in a sample of 1,078 pairs of fathers and sons.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8.2 Least Squares Method

The **least squares method** is a procedure for using sample data to find the estimated linear regression equation. To illustrate the least squares method, suppose data were collected from a sample of 10 Butler Trucking Company driving assignments. For the ith observation or driving assignment in the sample, x_i is the miles traveled and y_i is the travel time (in hours). The values of x_i and y_i for the 10 driving assignments in the sample are summarized in Table 8.1. We see that driving assignment 1, with $x_1 = 100$ and $y_1 = 9.3$, is a driving assignment of 100 miles and a travel time of 9.3 hours. Driving assignment 2, with $x_2 = 50$ and $y_2 = 4.8$, is a driving assignment of 50 miles and a travel time of 4.8 hours. The shortest travel time is for driving assignment 5, which requires 50 miles with a travel time of 4.2 hours.

Figure 8.3 is a scatter chart of the data in Table 8.1. Miles traveled is shown on the horizontal axis, and travel time (in hours) is shown on the vertical axis. Scatter charts for a regression task are constructed with the independent variable x on the horizontal axis and the dependent variable y on the vertical axis. The scatter chart enables us to observe the data graphically and to draw preliminary conclusions about the possible relationship between the variables.

What preliminary conclusions can be drawn from Figure 8.3? Longer travel times appear to coincide with more miles traveled. In addition, for these data, the relationship between the travel time and miles traveled appears to be approximated by a straight line; indeed, a positive linear relationship is indicated between x and y. We therefore choose the simple linear regression model to represent this relationship. Given that choice, our next task is to use the sample data in Table 8.1 to determine the values of b_0 and b_1 in the estimated simple linear regression equation. For the ith driving assignment, the estimated linear regression equation provides:

$$\hat{y}_i = b_0 + b_1 x_i \tag{8.3}$$

where

$\hat{y}_i$ = predicted travel time (in hours) for the ith driving assignment
b_0 = the y-intercept of the estimated regression line
b_1 = the slope of the estimated regression line
x_i = miles traveled for the ith driving assignment

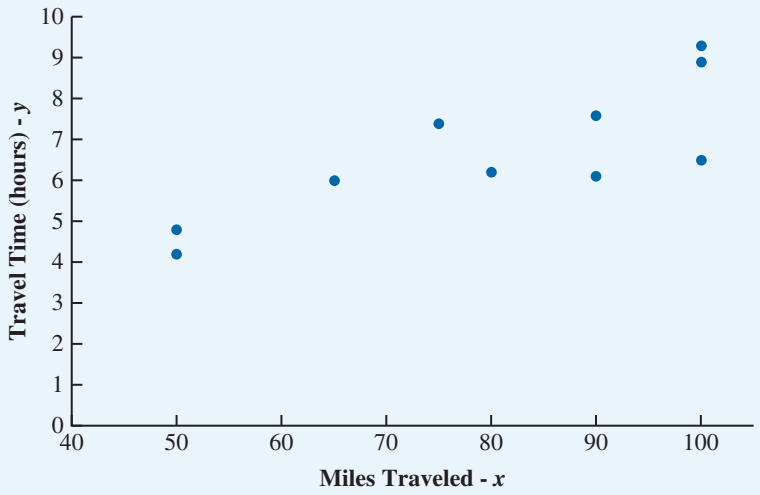
DATA*file*
butler

Table 8.1 Miles Traveled and Travel Time for 10 Butler Trucking Company Driving Assignments

Driving Assignment i	x = Miles Traveled	y = Travel Time (hours)
1	100	9.3
2	50	4.8
3	50	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.3 Scatter Chart of Miles Traveled and Travel Time for Sample of 10 Butler Trucking Company Driving Assignments



With y_i denoting the observed (actual) travel time for driving assignment i and $\hat{y}_i$ in equation (8.3) representing the predicted travel time for driving assignment i, every driving assignment in the sample will have an observed travel time y_i and a predicted travel time $\hat{y}_i$. For the estimated regression line to provide a good fit to the data, the differences between the observed travel times y_i and the predicted travel times $\hat{y}_i$ should be small.

The least squares method uses the sample data to provide the values of b_0 and b_1 that minimize the sum of the squares of the deviations between the observed values of the dependent variable y_i and the predicted values of the dependent variable $\hat{y}_i$. The criterion for the least squares method is given by equation (8.4).

Least Squares Equation

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_1 - b_0 - b_1 x_1)^2 \tag{8.4}$$

where

y_i = observed value of the dependent variable for the ith observation
$\hat{y}_i$ = predicted value of the dependent variable for the ith observation
n = total number of observations

The error we make using the linear regression model to estimate the mean value of the dependent variable for the ith observation is often written as $e_i = y_i - \hat{y}_i$ and is referred to as the ith **residual**. Using this notation, equation (8.4) can be rewritten as

$$\min \sum_{i=1}^{n} e_i^2$$

and we say that we are estimating the linear regression equation that minimizes the sum of squared errors.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Least Squares Estimates of the Simple Linear Regression Parameters

Although the values of b_0 and b_1 that minimize equation (8.4) can be calculated manually with equations (see note at end of this section), computer software such as Excel is generally used to calculate b_1 and b_0. For the Butler Trucking Company data in Table 8.1, an estimated slope of $b_1 = 0.0678$ and a y-intercept of $b_0 = 1.2739$ minimize the sum of squared errors (in the next section we show how to use Excel to obtain these values). Thus, our estimated simple linear regression equation is $\hat{y} = 1.2739 + 0.0678x_1$.

We interpret b_1 and b_0 as we would the slope and y-intercept of any straight line. The slope b_1 is the estimated change in the mean of the dependent variable y that is associated with a one-unit increase in the independent variable x. For the Butler Trucking Company model, we therefore estimate that, if the length of a driving assignment were 1 mile longer, the mean travel time for that driving assignment would be 0.0678 hour (or approximately 4 minutes) longer. The y-intercept b_0 is the estimated value of the dependent variable y when the independent variable x is equal to 0. For the Butler Trucking Company model, we estimate that if the driving distance for a driving assignment was 0 units (0 miles), the mean travel time would be 1.2739 units (1.2739 hours, or approximately 76 minutes). Can we find a plausible explanation for this? Perhaps the 76 minutes represent the time needed to prepare, load, and unload the vehicle, which is required for all trips regardless of distance and which therefore does not depend on the distance traveled. However, we cautiously note that to estimate the travel time for a driving distance of 0 miles, we have to extend the relationship we have found with simple linear regression well beyond the range of values for driving distance in our sample. Those sample values range from 50 to 100 miles, and this range represents the only values of driving distance for which we have empirical evidence of the relationship between driving distance and our estimated travel time.

It is important to note that the linear regression model is valid only over the **experimental region**, which is the range of values of the independent variables in the data used to estimate the model. Prediction of the value of the dependent variable outside the experimental region is called **extrapolation** and is risky. Because we have no empirical evidence that the relationship between y and x holds true for x values outside the range of x values in the data used to estimate the relationship, extrapolation is risky and should be avoided if possible. For Butler Trucking, this means that any prediction of the travel time for a driving distance less than 50 miles or greater than 100 miles is not a reliable estimate. Thus, any interpretation of β_0 based on the Butler Trucking Company data is unreliable and likely meaningless. However, if the experimental region for a linear regression model includes zero, the y-intercept will have a meaningful interpretation.

The estimated value of the y-intercept often results from extrapolation.

We can use the estimated linear regression equation and our known values for miles traveled for a driving assignment (x) to estimate mean travel time in hours. For example, the first driving assignment in Table 8.1 has a value for miles traveled of $x = 100$. We estimate the mean travel time in hours for this driving assignment to be

$$\hat{y}_i = 1.2739 + 0.0678(100) = 8.0539$$

The point estimate ŷ provided by the linear regression equation does not give us any information about the precision associated with the prediction. For that we must develop an interval estimate around the point estimate. In the last section of this chapter, we discuss the construction of interval estimates around the point predictions provided by a linear regression equation.

Since the travel time for this driving assignment was 9.3 hours, this linear regression estimate would have resulted in a residual of

$$e_1 = y_1 - \hat{y}_i = 9.3 - 8.0539 = 1.2461$$

The simple linear regression model underestimated travel time for this driving assignment by 1.2461 hours (approximately 74 minutes). Table 8.2 shows the predicted mean travel times, the residuals, and the squared residuals for all 10 driving assignments in the sample data. Note the following in Table 8.2:

- The sum of predicted values $\hat{y}_i$ is equal to the sum of the values of the dependent variable y.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

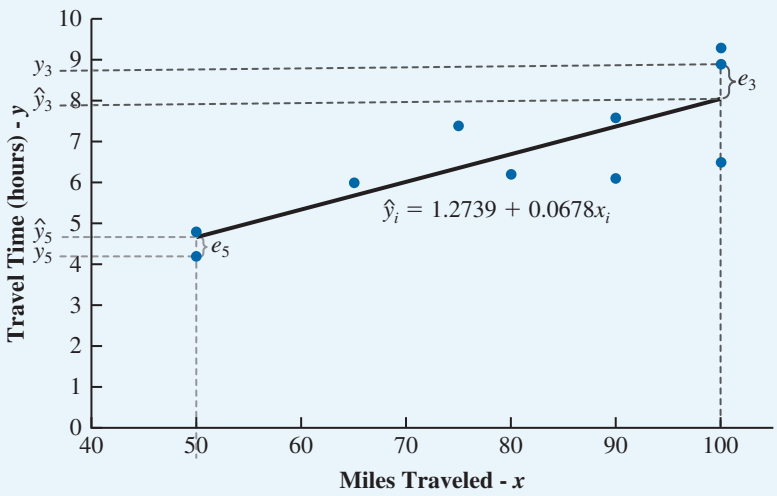
Table 8.2 Predicted Travel Time and Residuals for 10 Butler Trucking Company Driving Assignments

Driving Assignment i	x = Miles Traveled	y = Travel Time (hours)	$\hat{y}_i = b_0 + b_1 x_i$	$e_i = y_i - \hat{y}_i$	e_i^2
1	100	9.3	8.0565	1.2435	1.5463
2	50	4.8	4.6652	0.1348	0.0182
3	100	8.9	8.0565	0.8435	0.7115
4	100	6.5	8.0565	−1.5565	2.4227
5	50	4.2	4.6652	−0.4652	0.2164
6	80	6.2	6.7000	−0.5000	0.2500
7	75	7.4	6.3609	1.0391	1.0797
8	65	6.0	5.6826	0.3174	0.1007
9	90	7.6	7.3783	0.2217	0.0492
10	90	6.1	7.3783	−1.2783	1.6341
	Totals	67.0	67.0000	0.0000	8.0288

- The sum of the residuals e_i is 0.
- The sum of the squared residuals e_i^2 has been minimized.

These three points will always be true for a simple linear regression that is determined by equation (8.5). Figure 8.4 shows the simple linear regression line $\hat{y}_i = 1.2739 + 0.0678x_i$ superimposed on the scatter chart for the Butler Trucking Company data in Table 8.1. Figure 8.4 highlights the residuals for driving assignment 3 and driving assignment 5.

Figure 8.4 Scatter Chart of Miles Traveled and Travel Time for Butler Trucking Company Driving Assignments with Regression Line Superimposed



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The linear regression model underpredicts travel time for some driving assignments ($e_3 > 0$) and overpredicts travel time for others ($e_5 < 0$), but in general appears to fit the data relatively well.

In Figure 8.5, a vertical line is drawn from each point in the scatter chart to the linear regression line. Each of these vertical lines represents the difference between the actual driving time and the driving time we predict using linear regression for one of the assignments in our data. The length of each vertical line is equal to the absolute value of the residual for one of the driving assignments. When we square a residual, the resulting value is equal to the area of the square with the length of each side equal to the absolute value of the residual. In other words, the square of the residual for driving assignment 4, ($e_4 = (-1.5565)^2 = 2.4227$), is the area of a square for which the length of each side is 1.5565. Thus, when we find the linear regression model that minimizes the sum of squared errors for the Butler Trucking example, we are positioning the regression line in the manner that minimizes the sum of the areas of the 10 squares in Figure 8.5.

Using Excel's Chart Tools to Compute the Estimated Simple Linear Regression Equation

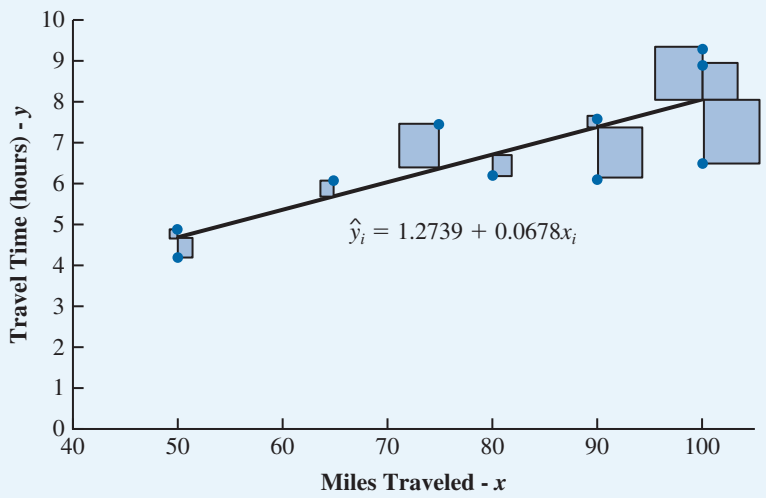
We can use Excel's chart tools to compute the estimated linear regression equation on a scatter chart of the Butler Trucking Company data in Table 8.1. After constructing a scatter chart (as shown in Figure 8.3) with Excel's chart tools, the following steps describe how to compute the estimated linear regression equation using the data in the worksheet:

Step 1. Right-click on any data point in the scatter chart and select **Add Trendline**

Step 2. When the **Format Trendline** task pane appears:
Select **Linear** in the **Trendline Options** area
Select **Display Equation on chart** in the **Trendline Options** area

The worksheet displayed in Figure 8.6 shows the original data, scatter chart, estimated regression line, and estimated linear regression equation.

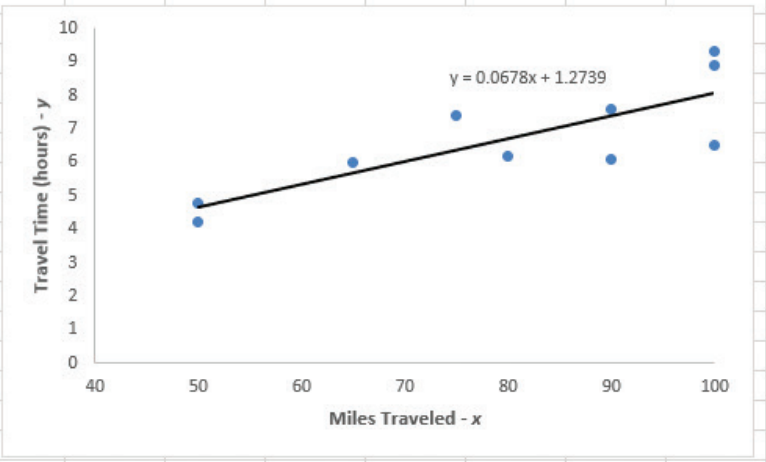
Figure 8.5 A Geometric Interpretation of the Least Squares Method



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.6 Scatter Chart and Estimated Regression Line for Butler Trucking Company

	A	B	C
1	Assignment	Miles	Time
2	1	100	9.3
3	2	50	4.8
4	3	100	8.9
5	4	100	6.5
6	5	50	4.2
7	6	80	6.2
8	7	75	7.4
9	8	65	6.0
10	9	90	7.6
11	10	90	6.1



Notes + Comments

1. Differential calculus can be used to show that the values of b_0 and b_1 that minimize expression (8.5) are given by:

Slope Equation

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

***y*-Intercept Equation**

$$b_0 = \bar{y} - b_1\bar{x}$$

where

x_i = value of the independent variable for the ith observation
y_i = value of the dependent variable for the ith observation
$\bar{x}$ = mean value for the independent variable
$\bar{y}$ = mean value for the dependent variable
n = total number of observations

2. Equation (8.4) minimizes the sum of the squared deviations between the observed values of the dependent variable y_i and the predicted values of the dependent variable $\hat{y}_1$. One alternative is to simply minimize the sum of the deviations between the observed values of the dependent variable y_i and the predicted values of the dependent variable $\hat{y}_i$. This is not a viable option because then negative deviations (observations for which the linear regression forecast exceeds the actual value) and positive deviations (observations for which the linear regression forecast is less than the actual value) offset each other. Another alternative is to minimize the sum of the absolute value of the deviations between the observed values of the dependent variable y_i and the predicted values of the dependent variable $\hat{y}_i$. It is possible to compute estimated linear regression parameters that minimize this sum of the absolute value of the deviations, but this approach is more difficult than the least squares approach.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8.3 Assessing the Fit of the Simple Linear Regression Model

For the Butler Trucking Company example, we developed the estimated linear regression equation $\hat{y}_i = 1.2739 + 0.0678x_i$ to approximate the linear relationship between the miles traveled (x) and travel time in hours (y). We now wish to assess how well the estimated linear regression equation fits the sample data. We begin by developing the intermediate calculations, referred to as the sums of squares.

The Sums of Squares

Recall that we found our estimated linear regression equation for the Butler Trucking Company example by minimizing the sum of squares of the residuals. This quantity, also known as the *sum of squares due to error*, is denoted by SSE.

Sum of Squares Due to Error

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8.5}$$

The value of SSE is a measure of the error (in squared units of the dependent variable) that results from using the estimated linear regression equation to predict the values of the dependent variable in the sample.

We have already shown the calculations required to compute the sum of squares due to error for the Butler Trucking Company example in Table 8.2. The squared residual or error for each observation in the data is shown in the last column of that table. After computing and squaring the residuals for each driving assignment in the sample, we sum them to obtain SSE = 8.0288 hours^2. Thus, SSE = 8.0288 measures the error in using the estimated linear regression equation $\hat{y}_i = 1.2739 + 0.0678x_i$ to predict travel time for the driving assignments in the sample.

Now suppose we are asked to predict travel time in hours without knowing the miles traveled for a driving assignment. Without knowledge of any related variables, we would use the sample mean $\bar{y}$ as a predictor of travel time for any given driving assignment. To find $\bar{y}$, we divide the sum of the actual driving times y_i from Table 8.2 (67) by the number of observations n in the data (10); this yields $\bar{y} = 6.7$.

Figure 8.7 provides insight on how well we would predict the values of y_i in the Butler Trucking Company example using $\bar{y} = 6.7$. From this figure, which again highlights the residuals for driving assignments 3 and 5, we can see that $\bar{y}$ tends to overpredict travel times for driving assignments that have relatively small values for miles traveled (such as driving assignment 5) and tends to underpredict travel times for driving assignments that have relatively large values for miles traveled (such as driving assignment 3).

In Table 8.3 we show the sum of squared deviations obtained by using the sample mean $\bar{y} = 6.7$ to predict the value of travel time in hours for each driving assignment in the sample. For the ith driving assignment in the sample, the difference $y_i - \bar{y}$ provides a measure of the error involved in using $\bar{y}$ to predict travel time for the ith driving assignment. The corresponding sum of squares, called the total sum of squares, is denoted by SST.

Total Sum of Squares, SST

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{8.6}$$

The sum at the bottom of the last column in Table 8.3 is the total sum of squares for Butler Trucking Company: SST = 23.9 hours^2.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.7 The Sample Mean $\bar{y}$ as a Predictor of Travel Time for Butler Trucking Company

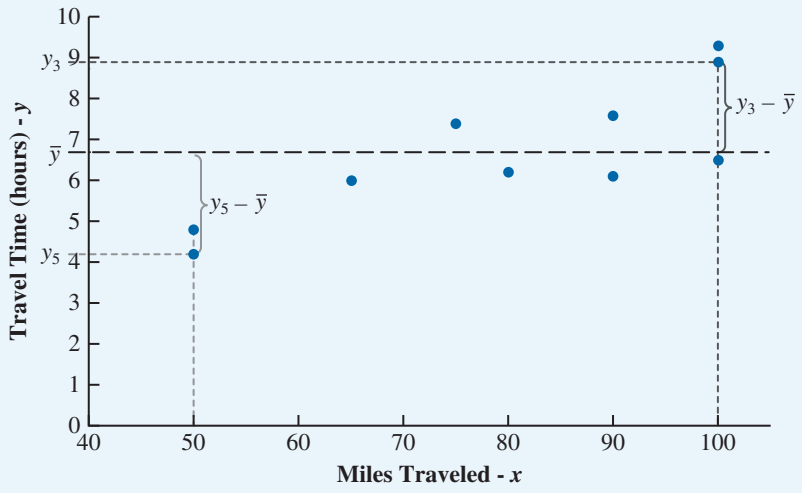


Table 8.3 Calculations for the Sum of Squares Total for the Butler Trucking Simple Linear Regression

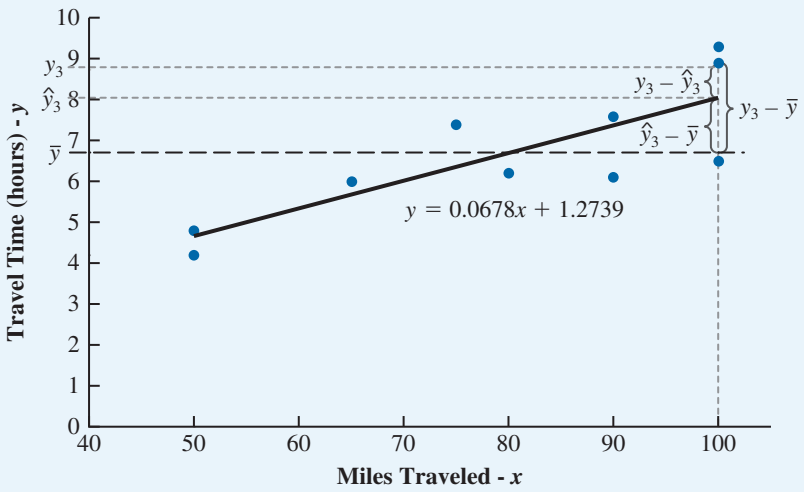
Driving Assignment i	x = Miles Traveled	y = Travel Time (hours)	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	100	9.3	2.6	6.76
2	50	4.8	−1.9	3.61
3	100	8.9	2.2	4.84
4	100	6.5	−0.2	0.04
5	50	4.2	−2.5	6.25
6	80	6.2	−0.5	0.25
7	75	7.4	0.7	0.49
8	65	6.0	−0.7	0.49
9	90	7.6	0.9	0.81
10	90	6.1	−0.6	0.36
	Totals	67.0	0	23.9

Now we put it all together. In Figure 8.8 we show the estimated regression line $\hat{y}_i = 1.2739 + 0.0678x_i$ and the line corresponding to $\bar{y} = 6.7$. Note that the points cluster more closely around the estimated regression line $\hat{y}_i = 1.2739 + 0.0678x_i$ than they do about the horizontal line $\bar{y} = 6.7$. For example, for the third driving assignment in the sample, we see that the error is much larger when $\bar{y} = 6.7$ is used to predict y_3 than when $\hat{y}_3 = 1.2739 + 0.0678(100) = 8.0539$ is used. We can think of SST as a measure of how well the observations cluster about the $\bar{y}$ line and SSE as a measure of how well the observations cluster about the $\hat{y}$ line.

To measure how much the $\hat{y}$ values on the estimated regression line deviate from $\bar{y}$, another sum of squares is computed. This sum of squares, called the *sum of squares due to regression*, is denoted by SSR.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.8 Deviations About the Estimated Regression Line and the Line $y = \bar{y}$ for the Third Butler Trucking Company Driving Assignment



Sum of Squares Due to Regression, SSR

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{8.7}$$

From the preceding discussion, we should expect that SST, SSR, and SSE are related. Indeed, the relationship among these three sums of squares is

$$\text{SST} = \text{SSR} + \text{SSE} \tag{8.8}$$

where

SST = total sum of squares
SSR = sum of squares due to regression
SSE = sum of squares due to error

The Coefficient of Determination

Now let us see how the three sums of squares, SST, SSR, and SSE, can be used to provide a measure of the goodness of fit for the estimated linear regression equation. The estimated linear regression equation would provide a perfect fit if every value of the dependent variable y_i happened to lie on the estimated regression line. In this case, $y_i - \hat{y}$ would be zero for each observation, resulting in SSE = 0. Because SST = SSR + SSE, we see that for a perfect fit SSR must equal SST, and the ratio (SSR/SST) must equal one. Poorer fits will result in larger values for SSE. Solving for SSE in equation (8.11), we see that SSE = SST − SSR. Hence, the largest value for SSE (and hence the poorest fit) occurs when SSR = 0 and SSE = SST. The ratio SSR/SST, which will take values between zero and one, is used to evaluate the goodness of fit for the estimated linear regression equation. This ratio is called the **coefficient of determination** and is denoted by r^2.

In simple linear regression, r^2 is often referred to as the simple coefficient of determination.

Coefficient of Determination

$$r^2 = \frac{\text{SSR}}{\text{SST}} \tag{8.9}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For the Butler Trucking Company example, the value of the coefficient of determination is

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{15.8712}{23.9} = 0.6641$$

The coefficient of determination r^2 is the square of the correlation between y_i and $\hat{y}_i$, and $0 \le r^2 \le 1$.

When we express the coefficient of determination as a percentage, r^2 can be interpreted as the percentage of the total sum of squares that can be explained by using the estimated linear regression equation. For Butler Trucking Company, we can conclude that 66.41% of the total sum of squares can be explained by using the estimated linear regression equation $\hat{y}_i = 1.2739 + 0.0678x_i$ to predict quarterly sales. In other words, 66.41% of the variability in the values of travel time in our sample can be explained by the linear relationship between the miles traveled and travel time.

Using Excel's Chart Tools to Compute the Coefficient of Determination

In Section 8.1 we used Excel's chart tools to construct a scatter chart and compute the estimated linear regression equation for the Butler Trucking Company data. We will now describe how to compute the coefficient of determination using the scatter chart in Figure 8.3.

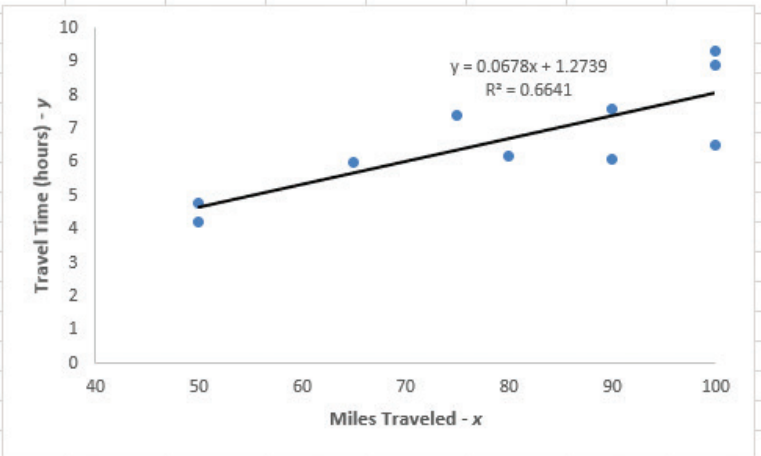
Note that Excel notates the coefficient of determination as R^2.

Step 1. Right-click on any data point in the scatter chart and select **Add Trendline**...
Step 2. When the **Format Trendline** task pane appears:
Select **Display R-squared value on chart** in the **Trendline Options** area

Figure 8.9 displays the scatter chart, the estimated linear regression equation, the estimated regression line, and the coefficient of determination for the Butler Trucking Company data. We see that $r^2 = 0.6641$.

Figure 8.9 Scatter Chart and Estimated Regression Line with Coefficient of Determination r^2 for Butler Trucking Company

	A	B	C
1	Assignment	Miles	Time
2	1	100	9.3
3	2	50	4.8
4	3	100	8.9
5	4	100	6.5
6	5	50	4.2
7	6	80	6.2
8	7	75	7.4
9	8	65	6.0
10	9	90	7.6
11	10	90	6.1



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8.4 The Multiple Linear Regression Model

We now extend our discussion to the study of how a dependent variable y is related to two or more independent variables. The concepts of a simple linear regression model and a simple linear regression equation introduced in the preceding sections are applicable in the multiple linear regression case. We will use q to denote the number of independent variables in the linear regression model. The equation that describes how the dependent variable y is related to the independent variables $x_1, x_2, \ldots, x_q$ and an error term is called the multiple linear regression model. We begin with the assumption that the multiple linear regression model takes the following form:

Multiple Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon \quad \textbf{(8.10)}$$

In the multiple linear regression model, $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$ are the parameters and the error term ϵ is a normally distributed random variable with a mean of zero and a constant variance across all observations. A close examination of this model reveals that y is a linear function of $x_1, x_2, \ldots, x_q$ plus the error term ϵ. As in simple linear regression, the error term accounts for the variability in y that cannot be explained by the linear effect of the q independent variables. The interpretation of the y-intercept β_0 in multiple linear regression is similar to the interpretation in simple linear regression; in a multiple linear regression model, β_0 is the mean of the dependent variable y when all of the independent variables $x_1, x_2, \ldots, x_q$ are equal to zero. On the other hand, the interpretation of the slope coefficients $\beta_1, \beta_2, \ldots, \beta_q$ in a multiple linear regression model differ in a subtle but important way from the interpretation of the slope β_1 in a simple linear regression model. In a multiple linear regression model the slope coefficient β_j represents the change in the mean value of the dependent variable y that corresponds to a one-unit increase in the independent variable x_j, *holding the values of all other independent variables in the model constant*. Thus, in a multiple linear regression model, the slope coefficient β_1 represents the change in the mean value of the dependent variable y that corresponds to a one-unit increase in the independent variable x_1, holding the values of $x_2, x_3, \ldots, x_q$ constant. Similarly, the slope coefficient β_2 represents the change in the mean value of the dependent variable y that corresponds to a one-unit increase in the independent variable x_2, holding the values of $x_1, x_3, \ldots, x_q$ constant.

Estimated Multiple Linear Regression Equation

In practice, the values of the population parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$ are not known and so must be estimated from sample data. A simple random sample is used to compute sample statistics $b_0, b_1, b_2, \ldots, b_q$ that are then used as the point estimators of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$. These sample statistics provide the following estimated multiple linear regression equation.

Estimated Multiple Linear Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_q x_q \quad \textbf{(8.11)}$$

where

$b_0, b_1, b_2, \ldots, b_q$ = the point estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$

$\hat{y}$ = estimated mean value of y given values for $x_1, x_2 \ldots, x_q$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Least Squares Method and Multiple Linear Regression

As with simple linear regression, in multiple linear regression we wish to find a model that results in small errors over the sample data. We continue to use the least squares method to develop the estimated multiple linear regression equation; that is, we find $b_0, b_1, b_2, \ldots, b_q$ that minimize the sum of squared residuals (the squared deviations between the observed values of the dependent variable y_i and the estimated values of the dependent variable $\hat{y}$):

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - b_0 - b_1 x_1 - \cdots - b_q x_q)^2 = \min \sum_{i=1}^{n} e_i^2 \qquad \textbf{(8.12)}$$

The estimation process for multiple linear regression is shown in Figure 8.10.

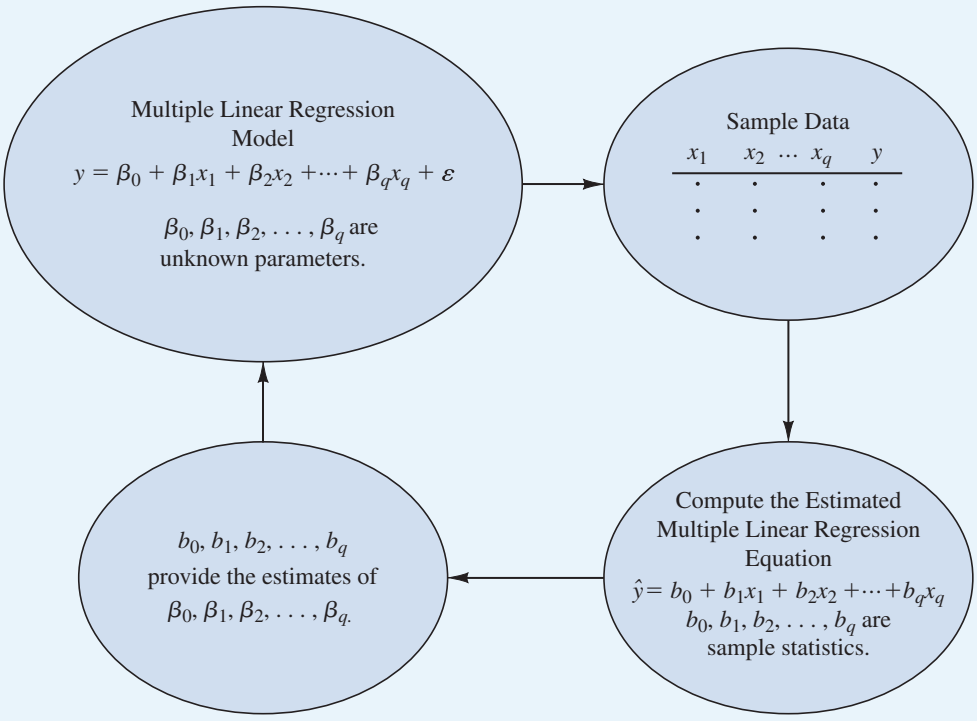
The estimated values of the dependent variable y are computed by substituting values of the independent variables $x_1, x_2, \ldots, x_q$ into the estimated multiple linear regression equation (8.11).

As in simple linear regression, it is possible to derive formulas that determine the values of the coefficients that minimize equation (8.12). However, these formulas involve the use of matrix algebra and are outside the scope of this text. Therefore, in presenting multiple linear regression, we focus on how computer software packages can be used to obtain the estimated linear regression equation and other information. The emphasis will be on how to construct and interpret a linear regression model.

Butler Trucking Company and Multiple Linear Regression

Recall that a major portion of Butler Trucking Company's business involves deliveries throughout its local area and that the managers want to estimate the total daily travel

Figure 8.10 The Estimation Process for Multiple Linear Regression



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

time for their drivers in order to develop better work schedules for the company's drivers.

Initially, the managers believed that the total daily travel time would be closely related to the number of miles traveled in making the daily deliveries. Based on a simple random sample of 10 driving assignments, we explored the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ to describe the relationship between travel time (y) and number of miles (x). As Figure 8.9 shows, we found that the estimated simple linear regression equation for our sample data is $\hat{y}_i = 1.2739 + 0.0678x_i$. With a coefficient of determination $r^2 = 0.6641$, the linear effect of the number of miles traveled explains 66.41% of the variability in travel time in the sample data, and so 33.59% of the variability in sample travel times remains unexplained. This result suggests to Butler's managers that other factors may contribute to the travel times for driving assignments. The managers might want to consider adding one or more independent variables to the model to explain some of the remaining variability in the dependent variable.

DATA*file*
butlerwithdeliveries

In considering other independent variables for their model, the managers felt that the number of deliveries made on a driving assignment also contributes to the total travel time. To support the development of a multiple linear regression model that includes both the number of miles traveled and the number of deliveries, they augment their original data with information on the number of deliveries for the 10 driving assignments in the original data and they collect new observations over several ensuing weeks. The new data, which consist of 300 observations, are provided in the file *butlerwithdeliveries*. Note that we now refer to the independent variables miles traveled as x_1 and the number of deliveries as x_2.

In multiple linear regression, R^2 is often referred to as the multiple coefficient of determination.

Our multiple linear regression with two independent variables will take the form $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$. The SSE, SST, and SSR are again calculated using equations (8.5), (8.6), and (8.7), respectively. Thus, the coefficient of determination, which in multiple linear regression is denoted by R^2, is again calculated using equation (8.9). We will now use Excel's Regression tool to calculate the values of the estimates b_0, b_1, b_2, and R^2.

Using Excel's Regression Tool to Develop the Estimated Multiple Linear Regression Equation

The following steps describe how to use Excel's Regression tool to compute the estimated multiple linear regression equation using the data in the worksheet.

Step 1. Click the **Data** tab in the Ribbon
Step 2. Click **Data Analysis** in the **Analysis** group
Step 3. Select **Regression** from the list of **Analysis Tools** in the **Data Analysis** tools box (shown in Figure 8.11) and click **OK**
Step 4. When the **Regression** dialog box appears (as shown in Figure 8.12):
Enter *D1:D301* in the **Input Y Range**: box
Enter *B1:C301* in the **Input X Range**: box
Select **Labels**
Select **Confidence Level**:
Enter *99* in the **Confidence Level**: box
Select **New Worksheet Ply**:
Click **OK**

Selecting **New Worksheet Ply***: tells Excel to place the linear regression output in a new worksheet. In the adjacent box, you can specify the name of the worksheet where the output is to be placed, or you can leave this blank and allow Excel to create a new worksheet to use as the destination for the linear regression results (as we are doing here).*

In the Excel output shown in Figure 8.13, the label for the independent variable x_1 is "Miles" (see cell A18), and the label for the independent variable x_2 is "Deliveries" (see cell A19). The estimated multiple linear regression equation, rounding the sample coefficients to four decimal places, is

$$\hat{y} = 0.1273 = 0.0672x_1 + 0.6900x_2 \quad \textbf{(8.13)}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If **Data Analysis** *does not appear in the* **Analysis** *group in the* **Data** *tab, you will have to load the Analysis ToolPak add-in into Excel. To do so, click the* **File** *tab in the Ribbon, and click* **Options**. *When the* **Excel Options** *dialog box appears, click* **Add-Ins** *from the menu. Next to* **Manage**:, *select* **Excel Add-ins**, *and click* **Go**. . . *at the bottom of the dialog box. When the* **Add-Ins** *dialog box appears, check the box next to* **Analysis ToolPak** *and click* **OK**.

Figure 8.11 Data Analysis Tools Box

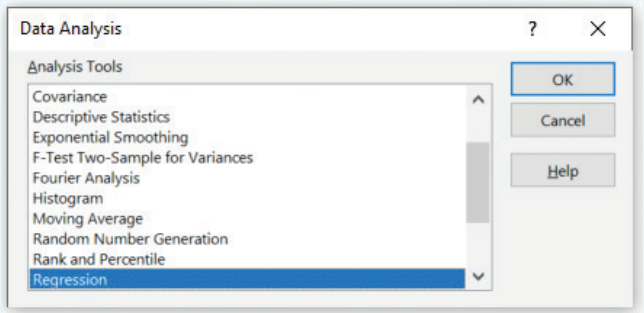
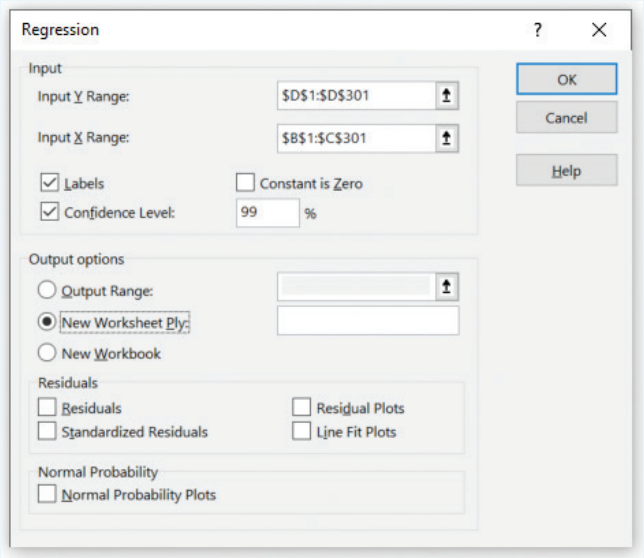


Figure 8.12 Regression Dialog Box



We interpret this model in the following manner:

- For a fixed number of deliveries, we estimate that the mean travel time will increase by 0.0672 hour (about 4 minutes) when the distance traveled increases by 1 mile.
- For a fixed distance traveled, we estimate that the mean travel time will increase by 0.69 hour (about 41 minutes) when the number of deliveries increases by 1 delivery.

The interpretation of the estimated y-intercept for this model (the expected mean travel time for a driving assignment with a distance traveled of 0 miles and no deliveries) is not meaningful because it is the result of extrapolation.

This model has a multiple coefficient of determination of $R^2 = 0.8173$. By adding the number of deliveries as an independent variable to our original simple linear regression, we now explain 81.73% of the variability in our sample values of the dependent variable, travel time. Because the simple linear regression with miles traveled as the sole independent variable explained 66.41% of the variability in our sample values of travel time, we can see that adding number of deliveries as an independent variable to our linear regression

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

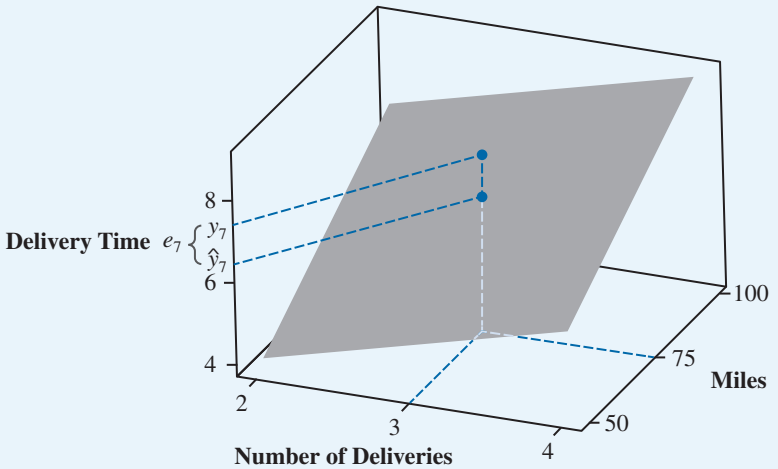
Figure 8.13 Excel Multiple Linear Regression Output for the Butler Trucking Company with Miles and Deliveries as Independent Variables

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.90407397							
5	R Square	0.817349743							
6	Adjusted R Square	0.816119775							
7	Standard Error	0.829967216							
8	Observations	300							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	915.5160626	457.7580313	664.5292419	2.2419E-110			
13	Residual	297	204.5871374	0.68884558					
14	Total	299	1120.1032						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	0.127337137	0.20520348	0.620540826	0.53537766	-0.276499931	0.531174204	-0.404649592	0.659323866
18	Miles	0.067181742	0.002454979	27.36551071	3.5398E-83	0.062350385	0.072013099	0.06081725	0.073546235
19	Deliveries	0.68999828	0.029521057	23.37308852	2.84826E-69	0.631901326	0.748095234	0.613465414	0.766531147

model resulted in explaining an additional 15.32% of the variability in our sample values of travel time. The addition of the number of deliveries to the model appears to have been worthwhile.

Using this multiple linear regression model, we now generate an estimated mean value of y for every combination of values of x_1 and x_2. Thus, instead of a regression line, we now have created a regression plane in three-dimensional space. Figure 8.14 provides the graph of the estimated regression plane for the Butler Trucking Company example and shows the seventh driving assignment in the data. Observe that as the plane slopes upward to larger

Figure 8.14 Graph of the Multiple Linear Regression Equation with Two Independent Variables



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

values of estimated mean travel time ($\hat{y}$) as either the number of miles traveled (x_1) or the number of deliveries (x_2) increases. Further, observe that the residual for a driving assignment when $x_1 = 75$ and $x_2 = 3$ is the difference between the observed y value and the estimated mean value of y given $x_1 = 75$ and $x_2 = 3$. Note that in Figure 8.14, the observed value lies above the regression plane, indicating that the linear regression model underestimates the expected driving time for the seventh driving assignment.

Notes + Comments

1. When using Excel's Regression tool, the data for the independent variables must be in adjacent columns or rows. Thus, you may have to rearrange the data in order to use Excel to run a particular linear multiple regression.
2. The sum of squares due to error, SSE, cannot become larger (and generally will become smaller) when independent variables are added to a linear regression model. Therefore, because SSR = SST − SSE, the SSR cannot become smaller (and generally becomes larger) when an independent variable is added to a linear regression model. Thus, R^2 = SSR / SST can never decrease as independent variables are added to the linear regression model.
3. Although we use linear regression analysis to estimate relationships between independent variables and the dependent variable, it does not provide information on whether these are cause-and-effect relationships. The analyst can conclude that a cause-and-effect relationship exists between an independent variable and a dependent variable only if there is a theoretical justification that the relationship is in fact causal. In the Butler Trucking Company multiple linear regression, we have found evidence of a relationship between distance traveled and travel time and evidence of a relationship between number of deliveries and travel time. Nonetheless, we cannot conclude from the linear regression model that changes in distance traveled x_1 cause changes in travel time y, and we cannot conclude that changes in number of deliveries x_2 cause changes in travel time y. The appropriateness of such cause-and-effect conclusions are left to supporting practical justification and to good judgment on the part of the analyst. Based on their practical experience, Butler Trucking's managers felt that increases in distance traveled and number of deliveries were likely causes of increased travel time. However, it is important to realize that the linear regression model itself provides no information about cause-and-effect relationships.

8.5 Inference and Linear Regression

The statistics $b_0, b_1, b_2, \ldots, b_q$ are point estimators of the population parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$; that is, each of these $q + 1$ estimates is a single value used as an estimate of the corresponding population parameter. Similarly, we use $\hat{y}$ as a point estimator of $E(y \mid x_1, x_2, \ldots, x_q)$, the conditional mean of y given values of $x_1, x_2, \ldots, x_q$.

However, we must recognize that samples do not replicate the population exactly. Different samples taken from the same population will result in different values of the point estimators $b_0, b_1, b_2, \ldots, b_q$; that is, the point estimators are random variables. If the values of a point estimator such as $b_0, b_1, b_2, \ldots, b_q$ change relatively little from sample to sample, the point estimator has low variability, and so the value of the point estimator that we calculate based on a random sample will likely be a reliable estimate of the population parameter. On the other hand, if the values of a point estimator change dramatically from sample to sample, the point estimator has high variability, and so the value of the point estimator that we calculate based on a random sample will likely be a less reliable estimate. How confident can we be in the estimates b_0, b_1, and b_2 that we developed for the Butler Trucking multiple linear regression model? Do these estimates have little variation and so are relatively reliable, or do they have so much variation that they have little meaning? We address the variability in potential values of the estimators through use of statistical inference.

Statistical inference is the process of making estimates and drawing conclusions about one or more characteristics of a population (the value of one or more parameters) through the analysis of sample data drawn from the population. In a linear regression model, we commonly use inference to estimate and draw conclusions about the following:

See Chapter 7 for a more thorough treatment of hypothesis testing and confidence intervals.

- The parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$.
- The mean value and/or the predicted value of the dependent variable y for specific values of the independent variables $x_1, x_2, \ldots, x_q$.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In our discussion of inference and linear regression, we will consider both **hypothesis testing** and **interval estimation**.

Conditions Necessary for Valid Inference in the Least Squares Linear Regression Model

It is assumed that the independent variables are linearly related to the dependent variable in the multiple linear regression model, which is expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon$$

The least squares method is used to develop values for $b_0, b_1, b_2, \ldots, b_q$, the estimates of the model parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, respectively. The resulting estimated multiple linear regression equation is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_q x_q$$

Although inference can provide greater understanding of the nature of relationships estimated through linear regression analysis, our inferences are valid only if the error term ε behaves in a certain way. Specifically, the validity of inferences in linear regression analysis depends on how well the following two conditions about the error term ε are met:

1. For any given combination of values of the independent variables $x_1, x_2, \ldots, x_q$, the population of potential error terms ϵ is normally distributed with a mean of 0 and a constant variance.
2. The values of ϵ are statistically independent.

The practical implication of normally distributed errors with a mean of zero and a constant variation for any given combination of values of $x_1, x_2, \ldots, x_q$ is that the parameter estimates are unbiased (i.e., they do not tend to over- or underpredict), possess consistent accuracy, and tend to err in small amounts rather than in large amounts. This first condition must be met for statistical inference in linear regression to be valid. The second condition is generally a concern when we collect data from a single entity over several periods of time and must also be met for statistical inference in linear regression to be valid in these instances. However, inferences in linear regression are generally reliable unless there are marked violations of these conditions.

Figure 8.15 illustrates these model conditions and their implications for a simple linear regression; note that in this graphical interpretation, the value of $E(y|x)$ changes linearly according to the specific value of x considered, and so the mean error is zero at each value of x. However, regardless of the x value, the error term ϵ and hence the dependent variable y are normally distributed, each with the same variance.

To evaluate whether the error of an estimated linear regression equation reasonably meets the two conditions, the sample residuals ($e_i = y_i - \hat{y}_i$ for observations $i = 1, \ldots, n$) need to be analyzed. There are many sophisticated diagnostic procedures for detecting whether the sample errors violate these conditions, but simple scatter charts of the residuals versus the predicted values of the dependent variable and the residuals versus the independent variables are an extremely effective method for assessing whether these conditions are violated. We should review the scatter chart for patterns in the residuals indicating that one or more of the conditions have been violated. As Figure 8.16 illustrates, at any given value of the horizontal-axis variable in these residual scatter plots, the center of the residuals should be approximately zero, the spread in the errors should be similar to the spread in error for other values of the horizontal-axis variable, and the errors should be symmetrically distributed with values near zero occurring more frequently than values that differ greatly from zero. A pattern in the residuals such as this gives us little reason to doubt the validity of inferences made on the linear regression model that generated the residuals.

While the residuals in Figure 8.16 show no discernible pattern, the residuals in the four panels of Figure 8.17 show examples of distinct patterns, each of which suggests a violation of at least one of the linear regression model conditions. Figure 8.17 shows plots of residuals from four different linear regression models, each showing a different pattern. In panel (a), the variation in the residuals (e) increases as the value of the independent variable increases, suggesting

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.15 Illustration of the Conditions for Valid Inference in Linear Regression

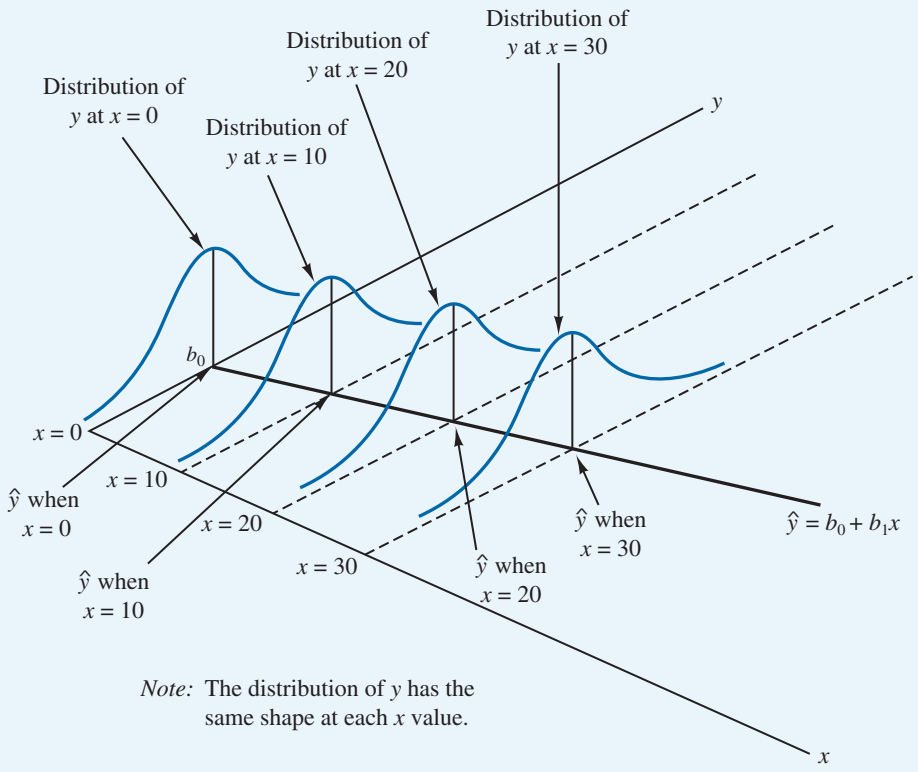
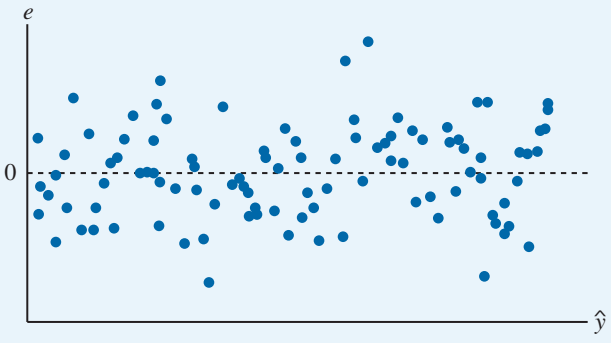


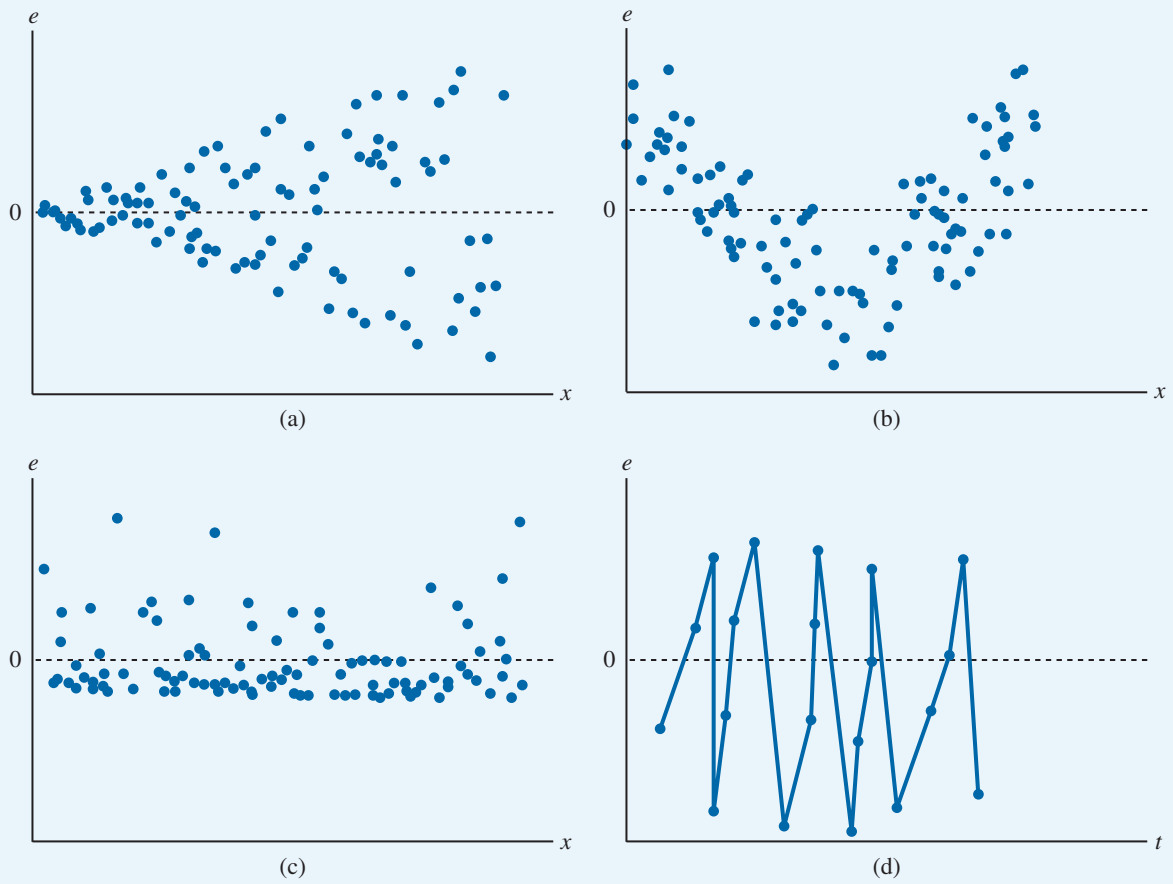
Figure 8.16 Example of a Random Error Pattern in a Scatter Chart of Residuals and Predicted Values of the Dependent Variable



that the residuals do not have a constant variance. In panel (b), the residuals are positive for small and large values of the independent variable but are negative for moderate values of the independent variable. This pattern suggests that the linear regression model underpredicts the value of the dependent variable for small and large values of the independent variable and overpredicts the value of the dependent variable for intermediate values of the independent

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.17 Examples of Diagnostic Scatter Charts of Residuals from Four Linear Regression Models



variable. In this case, the linear regression model does not adequately capture the relationship between the independent variable x and the dependent variable y. The residuals in panel (c) are not symmetrically distributed around 0; many of the negative residuals are relatively close to zero, while the relatively few positive residuals tend to be far from zero. This skewness suggests that the residuals are not normally distributed. Finally, the residuals in panel (d) are plotted over time t, which generally serves as an independent variable; that is, an observation is made at each of several (usually equally spaced) points in time. In this case, connected consecutive residuals allow us to see a distinct pattern across every set of four residuals; the second residual is consistently larger than the first and smaller than the third, whereas the fourth residual is consistently the smallest. This pattern, which occurs consistently over each set of four consecutive residuals in the chart in panel (d), suggests that the residuals generated by this model are not independent. A residual pattern such as this generally occurs when we have collected quarterly data and have not captured seasonal effects in the model. In each of these four instances, any inferences based on our linear regression model will likely not be reliable.

Frequently, the residuals do not meet these conditions either because an important independent variable has been omitted from the model or because the functional form of the model is inadequate to explain the relationships between the independent variables and the dependent variable. It is important to note that calculating the values of the estimates $b_0, b_1, b_2, \ldots, b_q$ does not require the errors to satisfy these conditions. However, the errors must satisfy these conditions in order for inferences (interval estimates for predicted values

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

of the dependent variable and confidence intervals and hypothesis tests of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$) to be reliable.

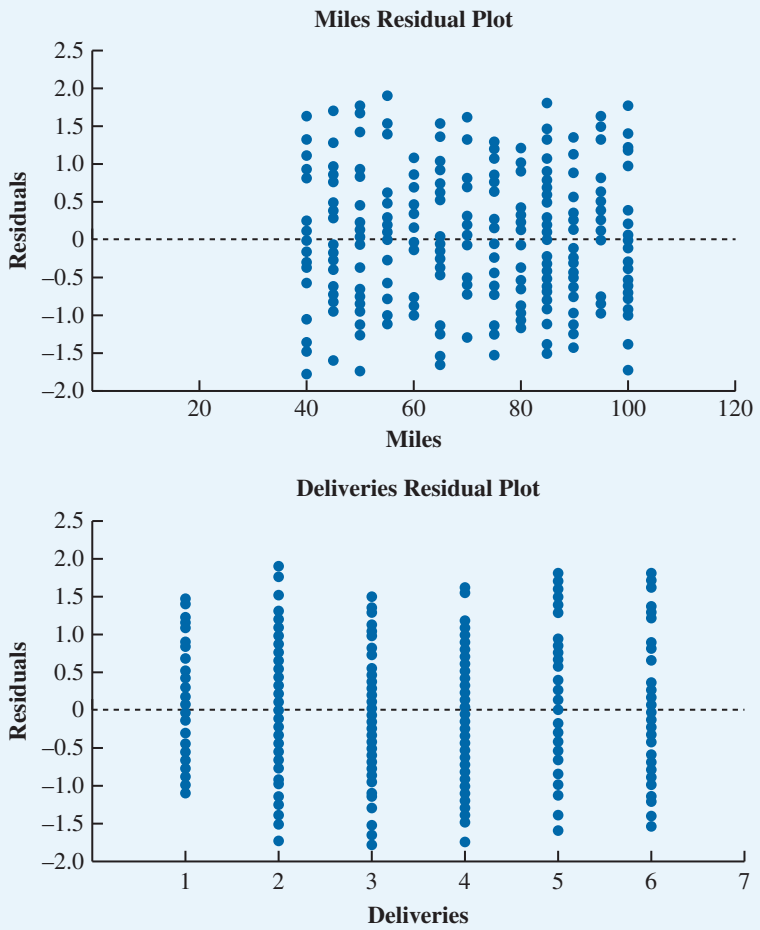
You can generate scatter charts of the residuals against each independent variable in the model when using Excel's Regression tool; to do so, select the **Residual Plots** option in the **Residuals** area of the **Regression** dialog box. Figure 8.18 shows residual plots produced by Excel for the Butler Trucking Company example for which the independent variables are miles (x_1) and deliveries (x_2).

The residuals at each value of miles appear to have a mean of zero, to have similar variances, and to be concentrated around zero. The residuals at each value of deliveries also appear to have a mean of zero, to have similar variances, and to be concentrated around zero. Although there appears to be a slight pattern in the residuals across values of deliveries, it is negligible and could conceivably be the result of random variation. Thus, this evidence provides little reason for concern over the validity of inferences about the linear regression model that we may perform.

A scatter chart of the residuals e against the predicted values of the dependent variables is also commonly used to assess whether the residuals of the linear regression model satisfy the conditions necessary for valid inference. To obtain the data to construct a scatter

Recall that in the Excel output shown in Figure 8.13, the label for the independent variable x_1 is "Miles" and the label for the independent variable x_2 is "Deliveries."

Figure 8.18 Excel Residual Plots for the Butler Trucking Company Multiple Linear Regression



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

chart of the residuals against the predicted values of the dependent variable using Excel's Regression tool, select the **Residuals** option in the **Residuals** area of the **Regression** dialog box (shown in Figure 8.12). This generates a table of predicted values of the dependent variable and residuals for the observations in the data; a partial list for the Butler Trucking multiple linear regression example is shown in Figure 8.19.

We can then use the Excel chart tool to create a scatter chart of these predicted values and residuals similar to the chart in Figure 8.20. The figure shows that the residuals at each predicted value of the dependent variable appear to have a mean of zero, to have similar variances, and to be concentrated around zero. Thus, the residuals provide little evidence that our linear regression model violates the conditions necessary for reliable inference. We can trust the inferences that we may wish to perform on our linear regression model.

Testing Individual Linear Regression Parameters

Once we ascertain that our linear regression model satisfies the conditions necessary for reliable inference reasonably well, we can begin testing hypotheses and building confidence intervals. Specifically, we may then wish to determine whether statistically significant relationships exist between the dependent variable y and each of the independent variables $x_1, x_2, \ldots, x_q$ individually. Note that if a β_j is zero, then the dependent variable y does not change when the independent variable x_j changes, and there is no linear relationship between y and x_j. Alternatively, if a β_j is not zero, there is a linear relationship between the dependent variable y and the independent variable x_j.

See Chapter 7 for a more in-depth discussion of hypothesis testing.

We use a ***t* test** to test the hypothesis that a parameter β_j is zero. The corresponding null and alternative hypotheses are as follows:

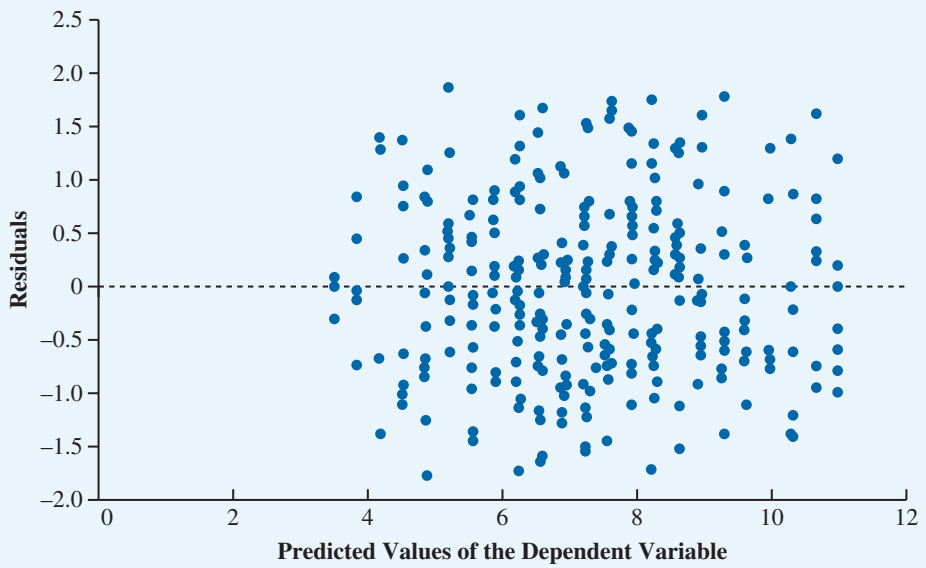
$$H_0: \beta_j = 0$$
$$H_a: \beta_j \neq 0$$

Figure 8.19 Table of the First Several Predicted Values $\hat{y}$ and Residuals e Generated by the Excel Regression Tool

23	RESIDUAL OUTPUT		
24			
25	*Observation*	*Predicted Time*	*Residuals*
26	1	9.605504464	-0.305504464
27	2	5.556419081	-0.756419081
28	3	9.605504464	-0.705504464
29	4	8.225507903	-1.725507903
30	5	4.8664208	-0.6664208
31	6	6.881873062	-0.681873062
32	7	7.235962632	0.164037368
33	8	7.254143492	-1.254143492
34	9	8.243688763	-0.643688763
35	10	7.553690482	-1.453690482
36	11	6.936415641	0.063584359
37	12	7.290505212	-0.290505212
38	13	9.287776613	0.312223387
39	14	5.874146931	0.625853069
40	15	6.954596501	0.245403499
41	16	5.556419081	0.443580919

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.20 Scatter Chart of Predicted Values $\hat{y}$ and Residuals e



The standard deviation of b_j is often referred to as the standard error of b_j. Thus, s_{b_j} provides an estimate of the standard error of b_j.

The test statistic for this t test is

$$t = \frac{b_j}{s_{b_j}} \tag{8.14}$$

where b_j is the point estimate of the parameter β_j and s_{b_j} is the estimated standard deviation of b_j.

As the value of b_j, the point estimate of β_j, deviates from zero in either direction, the evidence from our sample that the corresponding parameter β_j is not zero increases. Thus, as the magnitude of t increases (as t deviates from zero in either direction), we are more likely to reject the hypothesis that the parameter β_j is zero and so conclude that a relationship exists between the dependent variable y and the independent variable x_j.

Statistical software will generally report a p value for this test statistic; for a given value of t, this ***p* value** represents the probability of collecting a sample of the same size from the same population that yields a larger t statistic given that the value of β_j is actually zero. Thus, smaller p values indicate stronger evidence against the hypothesis that the value of β_j is zero (i.e., stronger evidence of a relationship between x_j and y). The hypothesis is rejected when the corresponding p value is smaller than some predetermined level of significance (usually 0.05 or 0.01).

The output of Excel's Regression tool provides the results of the t tests for each parameter. Refer again to Figure 8.13, which shows the multiple linear regression results for Butler Trucking with independent variables x_1 (labeled Miles) and x_2 (labeled Deliveries). The values of the parameter estimates b_0, b_1, and b_2 are located in cells B17, B18, and B19, respectively; the standard deviations s_{b_0}, s_{b_1}, and s_{b_2} are contained in cells C17, C18, and C19, respectively; the values of the t statistics for the hypothesis tests are in cells D17, D18, and D19, respectively; and the corresponding p values are in cells E17, E18, and E19, respectively.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Let's use these results to test the hypothesis that β_1 is zero. If we do not reject this hypothesis, we conclude that the mean value of y does not change when the value of x_1 changes, and so there is no relationship between driving time and miles traveled. We see in the Excel output in Figure 8.13 that the statistic for this test is 27.3655 and that the associated p value is very close to zero (3.5398E-83). This p value tells us that if the value of β_1 is actually zero, the probability we could collect a random sample of 300 observations from the population of Butler Trucking driving assignments that yields a t statistic with an absolute value greater than 27.3655 is practically zero. Such a small probability represents a highly unlikely scenario; thus, the small p value allows us to reject the hypothesis that $\beta_1 = 0$ for the Butler Trucking multiple linear regression example at a 0.01 level of significance or even at a far smaller level of significance. Thus, this data suggests that a relationship may exist between driving time and miles traveled.

Similarly, we can test the hypothesis that β_2 is zero. If we do not reject this hypothesis, we conclude that the mean value of y does not change when the value of x_2 changes, and so there is no relationship between driving time and number of deliveries. We see in the Excel output in Figure 8.13 that the t statistic for this test is 23.3731 and that the associated p value is very close to zero (2.84826E-69). This p value tells us that if the value of β_2 is actually zero, the probability we could collect a random sample of 300 observations from the population of Butler Trucking driving assignments that yields a t statistic with an absolute value greater than 23.3731 is practically zero. This is highly unlikely, and so the p value is sufficiently small to reject the hypothesis that $\beta_2 = 0$ for the Butler Trucking multiple linear regression example at a 0.01 level of significance or even at a far smaller level of significance. Thus, this data suggests that a relationship may exist between driving time and number of deliveries.

Finally, we can test the hypothesis that β_0 is zero in a similar fashion. If we do not reject this hypothesis, we conclude that the mean value of y is zero when the values of x_1 and x_2 are both zero, and so there is no driving time when a driving assignment is 0 miles and has 0 deliveries. We see in the Excel output that the t statistic for this test is 0.6205 and the associated p value is 0.5354. This p value tells us that if the value of β_0 is actually zero, the probability we could collect a random sample of 300 observations from the population of Butler Trucking driving assignments that yields a t statistic with an absolute value greater than 0.6205 is 0.5354. Thus, we do not reject the hypothesis that mean driving time is zero when a driving assignment is 0 miles and has 0 deliveries.

See Chapter 7 for a more in-depth discussion of confidence intervals.

We can also execute each of these hypothesis tests through confidence intervals. A **confidence interval** for a parameter β_i is an estimated interval believed to contain the true value of β_i at some level of confidence. The level of confidence, or **confidence level**, indicates how frequently interval estimates based on similar-sized samples from the same population using identical sampling techniques will contain the true value of β_i. Thus, when building a 95% confidence interval, we can expect that if we took similar-sized samples from the same population using identical sampling techniques, the corresponding interval estimates would contain the true value of β_i for 95% of the samples.

Although the confidence intervals for $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$ convey information about the variation in the estimates $b_0, b_1, b_2, \ldots, b_q$ that can be expected across repeated samples, they can also be used to test whether each of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$ is equal to zero in the following manner. To test that β_j is zero (i.e., there is no linear relationship between x_j and y) at some predetermined level of significance (say 0.05), first build a confidence interval at the $(1 - 0.05)100\%$ confidence level. If the resulting confidence interval does not contain zero, we conclude that β_j differs from zero at the predetermined level of significance.

The form of a confidence interval for β_j is as follows:

$$b_j \pm t_{a/2} s_{b_j}$$

where b_j is the point estimate of the parameter β_j, s_{b_j} is the estimated standard deviation of b_j, and $t_{\alpha/2}$ is a multiplier term based on the sample size and specified $100(1 - \alpha)\%$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

confidence level of the interval. More specifically, $t_{\alpha/2}$ is the t value that provides an area of $\alpha/2$ in the upper tail of a t distribution with $n - q - 1$ degrees of freedom.

Most software conducts linear regression analysis can also produce these confidence intervals. For example, the output of Excel's Regression tool for Butler Trucking, given in Figure 8.13, provides confidence intervals for β_1 (the slope coefficient associated with the independent variable x_1, labeled Miles) and β_2 (the slope coefficient associated with the independent variable x_2, labeled Deliveries), as well as the y-intercept β_0. The 95% confidence intervals for β_0, β_1, and β_2 are shown in cells F17:G17, F18:G18, and F19:G19, respectively. Neither of the 95% confidence intervals for β_1 and β_2 includes zero, so we can conclude that β_1 and β_2 each differ from zero at the 0.05 level of significance. On the other hand, the 95% confidence interval for β_0 does include zero, so we conclude that β_0 does not differ from zero at the 0.05 level of significance.

The Regression tool dialog box offers the user the opportunity to generate confidence intervals for β_0, β_1, and β_2 at a confidence level other than 95%. In this example, we chose to create 99% confidence intervals for β_0, β_1, and β_2, which in Figure 8.13 are given in cells H17:I17, H18:I18, and H19:I19, respectively. Neither of the 99% confidence intervals for β_1 and β_2 includes zero, so we can conclude that β_1 and β_2 each differs from zero at the 0.01 level of significance. On the other hand, the 99% confidence interval for β_0 does include zero, so we conclude that β_0 does not differ from zero at the 0.01 level of significance.

Addressing Nonsignificant Independent Variables

If we do not reject the hypothesis that β_j is zero, we conclude that there is no linear relationship between y and x_j. This leads to the question of how to handle the corresponding independent variable. Do we use the model as originally formulated with the nonsignificant independent variable, or do we reconstruct the linear regression model without the nonsignificant independent variable and use the new result? The approach to be taken depends on a number of factors, but ultimately whatever model we use should have a theoretical basis. If practical experience dictates that the nonsignificant independent variable has a relationship with the dependent variable, the independent variable should be left in the model. On the other hand, if the model sufficiently explains the dependent variable without the nonsignificant independent variable, then we should consider reconstructing the linear regression model without the nonsignificant independent variable. Note that it is possible that the estimates of the other variables' coefficients and their p values may change considerably when we remove the nonsignificant independent variable from the model.

The appropriate treatment of the inclusion or exclusion of the y-intercept when b_0 is not statistically significant may require special consideration. For example, in the Butler Trucking multiple linear regression model, recall that the p value for b_0 is 0.5354, suggesting that this estimate of β_0 is not statistically significant. Should we remove the y-intercept from this model because it is not statistically significant? Excel provides functionality to remove the y-intercept from the model by selecting **Constant is zero** in Excel's Regression tool. This will force the y-intercept to go through the origin (when the independent variables $x_1, x_2, ..., x_q$ all equal zero, the estimated value of the dependent variable will be zero). However, doing this can substantially alter the estimated slopes in the linear regression model and result in a model that yields less accurate predicted values of the dependent variable. The primary purpose of the linear regression model is to explain or predict values of the dependent variable corresponding to values of the independent variables within the experimental region. Therefore, it is generally advised that y-intercept of a linear regression model should not be forced to be zero. In a situation for which there are strong *a priori* reasons for believing that the dependent variable is equal to zero when the values of all independent variables in the model are equal to zero, it is better to collect data for which the values of the independent variables are at or near zero in order to allow the estimated linear regression equation to empirically validate this belief and avoid extrapolation. If data for which the values of the independent variables are at or near zero is not obtainable, and the linear regression model is intended to be used to explain or predict values of the dependent

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

variable at or near y-intercept, then forcing the y-intercept to be zero may be a necessary action, although it results in extrapolation and may result in poor predictions. A common business example of a linear regression model for which the y-intercept would be expected to be zero is when the output of a labor-intensive production process is the dependent variable and hours of labor is the independent variable. Because the production process is labor intense, we would expect no output when the value of labor hours is zero.

Multicollinearity

We use the term *independent variable* in regression analysis to refer to any variable used to predict or explain the value of the dependent variable. The term does not mean, however, that the independent variables themselves are independent from each other in any statistical sense. On the contrary, most independent variables in a multiple linear regression problem are correlated with one another to some degree. For example, in the Butler Trucking example involving the two independent variables x_1 (miles traveled) and x_2 (number of deliveries), we could compute the sample correlation coefficient r_{x_1,x_2} to determine the extent to which these two variables are related. Doing so yields $r_{x_1,x_2} = 0.16$. Thus, we find some degree of linear association between the two independent variables. In multiple linear regression analysis, **multicollinearity** refers to the correlation among the independent variables.

To gain a better perspective of the potential problems of multicollinearity, let us consider a modification of the Butler Trucking example. Instead of x_2 being the number of deliveries, let x_2 denote the number of gallons of gasoline consumed. Clearly, x_1 (the miles traveled) and x_2 are now related; that is, we know that the number of gallons of gasoline used depends to a large extent on the number of miles traveled. Hence, we would conclude logically that x_1 and x_2 are highly correlated independent variables and that multicollinearity is present in the model. The data for this example are provided in the file *butlerwithgasconsumption*.

DATA*file*
butlerwithgasconsumption

Using Excel's Regression tool, we obtain the results shown in Figure 8.21 for our multiple linear regression. When we conduct a t test to determine whether β_1 is equal to zero, the p value is very close to zero (3.1544E-07), and so we reject this hypothesis and conclude that travel time is related to miles traveled. On the other hand, when we conduct a t test to determine whether β_2 is equal to zero, we find a p value of 0.6588, and so we do not reject this hypothesis. Does this mean that travel time is not related to gasoline consumption? Not necessarily.

What it probably means in this instance is that, with x_1 already in the model, x_2 does not make a significant marginal contribution to predicting the value of y. This interpretation makes sense within the context of the Butler Trucking example; if we know the miles traveled, we do not gain much new information that would be useful in predicting driving time by also knowing the amount of gasoline consumed. We can see this in the scatter chart in Figure 8.22; miles traveled and gasoline consumed are strongly related.

If any estimated parameters $b_1, b_2, \ldots, b_q$ or associated p values change dramatically when a new independent variable is added to the model (or an existing independent variable is removed from the model), multicollinearity is likely present. Looking for changes such as these is sometimes used as a way to detect multicollinearity.

Even though we rejected the hypothesis that β_1 is equal to zero in the model corresponding to Figure 8.21, a comparison to Figure 8.13 shows the value of the t statistic is much smaller and the p value substantially larger than in the multiple linear regression model that includes miles driven and number of deliveries as the independent variables. The evidence against the hypothesis that β_1 is equal to zero is weaker in the multiple linear regression model that includes miles driven and gasoline consumed as the independent variables because of the high correlation between these two independent variables.

To summarize, in t tests for the significance of individual parameters, the difficulty caused by multicollinearity is that it is possible to conclude that a parameter associated with one of the multicollinear independent variables is not significantly different from zero when the independent variable actually has a strong relationship with the dependent variable. This problem is avoided when there is little correlation among the independent variables.

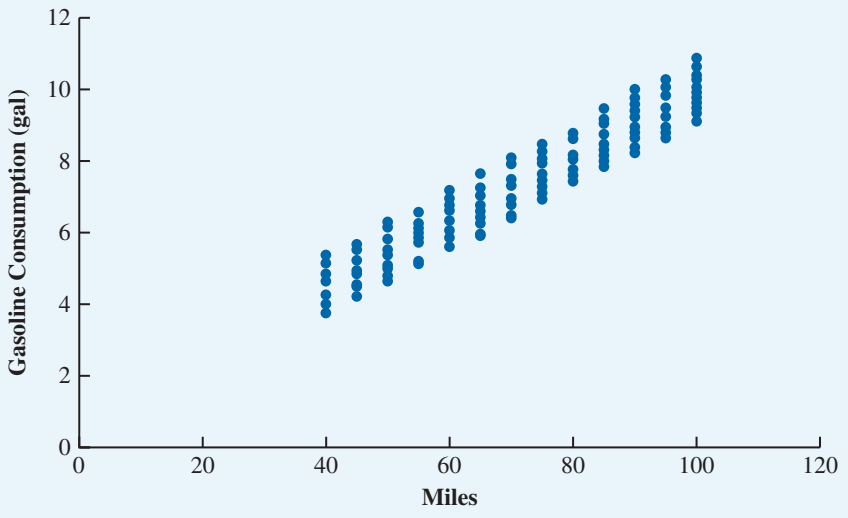
Statisticians have developed several tests for determining whether multicollinearity is strong enough to cause problems. In addition to the initial understanding of the nature

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.21 Excel Multiple Regression Output for the Butler Trucking Company with Miles and Gasoline Consumption as Independent Variables

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.69406354							
5	R Square	0.481724198							
6	Adjusted R Square	0.478234125							
7	Standard Error	1.398077545							
8	Observations	300							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	539.5808158	269.7904079	138.0269794	4.09542E-43			
13	Residual	297	580.5223842	1.954620822					
14	Total	299	1120.1032						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	2.493095385	0.33669895	7.404523781	1.36703E-12	1.830477398	3.155713373	1.620208758	3.365982013
18	Miles	0.074701825	0.014274552	5.233216928	3.15444E-07	0.046609743	0.102793908	0.037695279	0.111708371
19	Gasoline Consumption	-0.067506102	0.152707928	-0.442060235	0.658767336	-0.368032789	0.233020584	-0.463398955	0.328386751

Figure 8.22 Scatter Chart of Miles and Gasoline Consumed for Butler Trucking Company



of the relationships between the various pairs of variables that we can gain through scatter charts such as the chart shown in Figure 8.22, correlations between pairs of independent variables can be used to identify potential problems. According to a common rule-of-thumb test, multicollinearity is a potential problem if the absolute value of the sample correlation

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

See Chapter 2 for a more in-depth discussion of correlation and how to compute it with Excel.

coefficient exceeds 0.7 for any two of the independent variables. We can use the Excel function

$$=CORREL(B2:B301, C2:C301)$$

to find that the correlation between Miles (in column B) and Gasoline Consumed (in column C) in the file *butlerwithgasconsumption* is $r_{\text{Miles, Gasoline Consumed}} = 0.9572$, which supports the conclusion that Miles and Gasoline Consumed are multicollinear. Similarly, we can use the Excel function

$$=CORREL(B2:B301, D2:D301)$$

to show that the correlation between Miles (in column B) and Deliveries (in column D) for the sample data is $r_{\text{Miles, Deliveries}} = 0.0258$. This supports the conclusion that Miles and Deliveries are not multicollinear. Other tests for multicollinearity are more advanced and beyond the scope of this text.

The primary consequence of multicollinearity is that it increases the standard deviation of b_0, b_1, $b_2 \ldots, b_q$ and $\hat{y}$, and so inference based on these estimates is less precise than it should be. This means that confidence intervals for β_0, β_1, β_2, . . . , β_q and predicted values of the dependent variable are wider than they should be. Thus, we are less likely to reject the hypothesis that an individual parameter b_j is equal to zero than we otherwise would be, and multicollinearity leads us to conclude that an independent variable x_j is not related to the dependent variable y when they in fact are related. In addition, multicollinearity can result in confusing or misleading estimates $b_1, b_2, \ldots, b_q$. Therefore, if a primary objective of the linear regression analysis is inference, to explain the relationship between a dependent variable y and a set of independent variables $x_1, x_2 \ldots, x_q$, you should, if possible, avoid including independent variables that are highly correlated in the linear regression model. For example, when a pair of independent variables is highly correlated it is common to simply include only one of these independent variables in the linear regression model. When decision makers have reason to believe that substantial multicollinearity is present and they choose to retain the highly correlated independent variables in the model, they must realize that separating the relationships between each of the individual independent variables and the dependent variable is difficult (and maybe impossible). On the other hand, multicollinearity does not affect the predictive capability of a linear regression model, so if the primary objective is prediction or forecasting, then multicollinearity is not a concern.

Notes + Comments

1. In multiple linear regression we can test the null hypothesis that the parameters $\beta_1, \beta_2, \ldots, \beta_q$ are all equal to zero (H_0: $\beta_1 = \beta_2 = \cdots = \beta_q = 0$, H_a: at least one $\beta_j \neq 0$ for $j = 1, \ldots, q$) with an F test based on the F probability distribution. The test statistic generated by the sample data for this test is

$$F = \frac{SSR/q}{SSE/(n - q - 1)}$$

where SSR and SSE are as defined by equations (8.5) and (8.7), q is the number of independent variables in the linear regression model, and n is the number of observations in the sample. If the p value corresponding to the F statistic is smaller than some predetermined level of significance (usually 0.05 or 0.01), this leads us to reject the hypothesis that the values of $\beta_1, \beta_2, \ldots, \beta_q$ are all zero, and we would conclude that there is an overall relationship; otherwise, we conclude that there is no overall relationship.

The output of Excel's Regression tool provides the results of the F test; in Figure 8.13, which shows the multiple linear regression results for Butler Trucking with independent variables x_1 (labeled "Miles") and x_2 (labeled "Deliveries"), the value of the F statistic and the corresponding p value are in cells E12 and F12, respectively. From the Excel output in Figure 8.13 we see that

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the p value for the F test is essentially 0. Thus, the p value is sufficiently small to allow us to reject the hypothesis that no overall relationship exists at the 0.01 level of significance.

2. Finding a significant relationship between an independent variable x_j and a dependent variable y in a linear regression does not enable us to conclude that the relationship is linear. We can state only that x_j and y are related and that a linear relationship explains a statistically significant portion of the variability in y over the range of values for x_j observed in the sample.
3. Note that a review of the correlations of pairs of independent variables is not always sufficient to entirely uncover multicollinearity. The problem is that sometimes one independent variable is highly correlated with some combination of several other independent variables. If you suspect that one independent variable is highly correlated with a combination of several other independent variables, you can use multiple linear regression to assess whether the sample data support your suspicion. Suppose that your original linear regression model includes the independent variables x_1, $x_2, \ldots, x_q$ and that you suspect that x_1 is highly correlated with a subset of the other independent variables $x_2, \ldots, x_q$. Then construct the multiple linear regression for which x_1 is the dependent variable to be explained by the subset of the independent variables $x_2, \ldots, x_q$ that you suspect are highly correlated with x_1. The coefficient of determination R^2 for this linear regression model provides an estimate of the strength of the relationship between x_1 and the subset of the other independent variables $x_2, \ldots, x_q$ that you suspect are highly correlated with x_1. As a rule of thumb, if the coefficient of determination R^2 for this linear regression model exceeds 0.50, multicollinearity between x_1 and the subset of the other independent variables $x_2, \ldots, x_q$ is a concern.
4. When working with a small number of observations, assessing the conditions necessary for inference to be valid in linear regression can be extremely difficult. Similarly, when working with a small number of observations, assessing multicollinearity can also be difficult.
5. In some instances, the values of the independent variables to be used to estimate the value of dependent variable are not known. For example, a company may include its competitor's price as an independent variable in a linear regression model to be used to estimate demand for one of its products in some future period. It is unlikely that the competitor's price in some future period will be known by this company, and so the company may estimate what the competitor's price will be and substitute this estimated value into the linear regression equation.

   In such instances, estimated values of the independent variables are sometimes substituted into the linear regression equation to produce an estimated value of the dependent variable. The result can be useful, but one must proceed with caution as an inaccurate estimate of the value of any independent variable can create an inaccurate estimate of the dependent variable.

8.6 Categorical Independent Variables

Thus far, the examples we have considered have involved quantitative independent variables such as the miles traveled and the number of deliveries. In many situations, however, we must work with categorical independent variables such as marital status (married, single) and method of payment (cash, credit card, check). The purpose of this section is to show how categorical variables are handled in linear regression analysis. To illustrate the use and interpretation of a categorical independent variable, we will again consider the Butler Trucking Company example.

Butler Trucking Company and Rush Hour

Dummy variables are sometimes referred to as indicator variables.

Several of Butler Trucking's driving assignments require the driver to travel on a congested segment of a highway during the afternoon rush hour. Management believes that this factor may also contribute substantially to variability in the travel times across driving assignments. How do we incorporate information on which driving assignments include travel on a congested segment of a highway during the afternoon rush hour into a linear regression model?

The previous independent variables we have considered (such as the miles traveled and the number of deliveries) have been quantitative, but this new variable is categorical and will require us to define a new type of variable called a **dummy variable**.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To incorporate a variable that indicates whether a driving assignment included travel on this congested segment of a highway during the afternoon rush hour into a model that currently includes the miles traveled (x_1) and the number of deliveries (x_2), we define the following variable:

$$x_3 = \begin{cases} 0 \text{ if an assignment did not include travel on the congested segment of highway during afternoon rush hour} \\ 1 \text{ if an assignment included travel on the congested segment of highway during afternoon rush hour} \end{cases}$$

See Chapters 2 and 3 for step-by-step descriptions of how to construct charts in Excel.

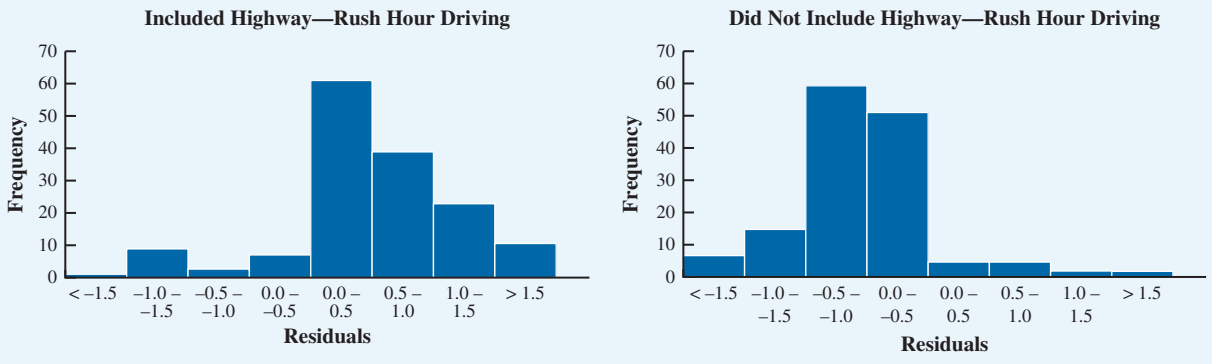
Will this dummy variable add valuable information to the current Butler Trucking linear regression model? A review of the residuals produced by the current model may help us make an initial assessment. Using Excel chart tools, we can create a frequency distribution and a histogram of the residuals for driving assignments that included travel on a congested segment of a highway during the afternoon rush hour period. We then create a frequency distribution and a histogram of the residuals for driving assignments that did not include travel on a congested segment of a highway during the afternoon rush hour period. The two histograms are shown in Figure 8.23.

Recall that the residual for the *i*th observation is $e_i = y_i - \hat{y}_i$, which is the difference between the observed and the predicted values of the dependent variable. The histograms in Figure 8.23 show that driving assignments that included travel on a congested segment of a highway during the afternoon rush hour period tend to have positive residuals, which means we are generally underpredicting the travel times for those driving assignments. Conversely, driving assignments that did not include travel on a congested segment of a highway during the afternoon rush hour period tend to have negative residuals, which means we are generally overpredicting the travel times for those driving assignments. These results suggest that the dummy variable could potentially explain a substantial proportion of the variance in travel time that is unexplained by the current model, and so we proceed by adding the dummy variable x_3 to the current Butler Trucking multiple linear regression model. Using Excel's Regression tool to develop the estimated linear regression equation on the data in the file *butlerhighway*, we obtain the Excel output in Figure 8.24. The estimated linear regression equation is

DATA*file*
butlerhighway

$$\hat{y} = -0.3302 + 0.0672x_1 + 0.6735x_2 + 0.9980x_3 \quad \textbf{(8.15)}$$

Figure 8.23 Histograms of the Residuals for Driving Assignments That Include Travel on a Congested Segment of a Highway During the Afternoon Rush Hour and Residuals for Driving Assignments That Did Not



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.24 Excel Data and Output for Butler Trucking with Miles Traveled (x_1), Number of Deliveries (x_2), and the Highway Rush Hour Dummy Variable (x_3) as the Independent Variables

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.940107228							
5	R Square	0.8838016							
6	Adjusted R Square	0.882623914							
7	Standard Error	0.663106426							
8	Observations	300							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	989.9490008	329.9830003	750.455757	5.7766E-138			
13	Residual	296	130.1541992	0.439710132					
14	Total	299	1120.1032						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	-0.330229304	0.167677925	-1.969426232	0.04983651	-0.66022126	-0.000237349	-0.764941128	0.104482519
18	Miles	0.067220302	0.00196142	34.27125147	4.7852E-105	0.063360208	0.071080397	0.062135243	0.072305362
19	Deliveries	0.67351584	0.023619993	28.51465081	6.74797E-87	0.627031441	0.720000239	0.612280051	0.734751629
20	Highway	0.9980033	0.076706582	13.0106605	6.49817E-31	0.847043924	1.148962677	0.799138374	1.196868226

Interpreting the Parameters

After checking to make sure the regression model satisfies the conditions for inference and does not suffer from serious multicollinearity, we can consider inference on our results. The p values for the t tests of miles traveled (p value = 4.7852E-105), number of deliveries (p value = 6.7480E-87), and the rush hour driving dummy variable (p value = 6.4982E-31) are all extremely small, indicating that each of these independent variables has a statistical relationship with travel time. The model estimates that the mean travel time of a driving assignment increases by:

- 0.0672 hour (about 4 minutes) for every increase of 1 mile traveled, holding constant the number of deliveries and whether the driving assignment route requires the driver to travel on the congested segment of a highway during the afternoon rush hour.
- 0.6735 hour (about 40 minutes) for every delivery, holding constant the number of miles traveled and whether the driving assignment route requires the driver to travel on the congested segment of a highway during the afternoon rush hour.
- 0.9980 hour (about 60 minutes) if the driving assignment route requires the driver to travel on the congested segment of a highway during the afternoon rush hour, holding constant the number of miles traveled and the number of deliveries.

In addition, $R^2 = 0.8838$ indicates that the linear regression model explains approximately 88.4% of the variability in travel time for the driving assignments in the sample. Thus, equation (8.15) should prove helpful in estimating the travel time necessary for the various driving assignments.

To understand how to interpret the linear regression model when a categorical variable is present, let's compare the model for the case when $x_3 = 0$ (the driving assignment does

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

not include travel on congested highways) and when $x_3 = 1$ (the driving assignment does include travel on congested highways). In the case that $x_3 = 0$, we have

$$\begin{aligned}\hat{y} &= -0.3302 + 0.0672x_1 + 0.6735x_2 + 0.9980(0) \\ &= -0.3302 + 0.0672x_1 + 0.6735x_2\end{aligned} \tag{8.16}$$

In the case that when $x_3 = 1$, we have

$$\begin{aligned}\hat{y} &= -0.3302 + 0.0672x_1 + 0.6735x_2 + 0.9980(1) \\ &= 0.6678 + 0.0672x_1 + 0.6735x_2\end{aligned} \tag{8.17}$$

Comparing equations (8.16) and (8.17), we see that the mean travel time has the same linear relationship with x_1 and x_2 for both driving assignments that include travel on the congested segment of highway during the afternoon rush hour period and driving assignments that do not. However, the y-intercept is -0.3302 in equation (8.16) and 0.6678 in equation (8.17). That is, 0.9980 is the difference between the mean travel time for driving assignments that include travel on the congested segment of highway during the afternoon rush hour and the mean travel time for driving assignments that do not.

In effect, the use of a dummy variable provides two estimated linear regression equations that can be used to predict the travel time: One that corresponds to driving assignments that include travel on the congested segment of highway during the afternoon rush hour period, and one that corresponds to driving assignments that do not include such travel.

More Complex Categorical Variables

The categorical variable for the Butler Trucking Company example had two levels: (1) driving assignments that include travel on the congested segment of highway during the afternoon rush hour and (2) driving assignments that do not. As a result, defining a dummy variable with a value of zero indicating a driving assignment that does not include travel on the congested segment of highway during the afternoon rush hour and a value of one indicating a driving assignment that includes such travel was sufficient. However, when a categorical variable has more than two levels, care must be taken in both defining and interpreting the dummy variables. As we will show, if a categorical variable has k levels, $k - 1$ dummy variables are required, with each dummy variable corresponding to one of the levels of the categorical variable and coded as 0 or 1.

For example, suppose a manufacturer of vending machines organized the sales territories for a particular state into three regions: A, B, and C. The managers want to use linear regression analysis to help predict the number of vending machines sold per week. With the number of units sold as the dependent variable, they are considering several independent variables (the number of sales personnel, advertising expenditures, etc.). Suppose the managers believe that sales region is also an important factor in predicting the number of units sold. Because sales region is a categorical variable with three levels (A, B, and C), we will need $3 - 1 = 2$ dummy variables to represent the sales region. Selecting Region A to be the "reference" region, each dummy variable can be coded 0 or 1 as follows:

$$x_1 = \begin{cases} 1 \text{ if sales Region B} \\ 0 \text{ otherwise} \end{cases} \qquad x_2 = \begin{cases} 1 \text{ if sales Region C} \\ 0 \text{ otherwise} \end{cases}$$

With this definition, we have the following values of x_1 and x_2:

Region	x_1	x_2
A	0	0
B	1	0
C	0	1

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The linear regression equation relating the estimated mean number of units sold to the dummy variables is written as

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Observations corresponding to Region A correspond to $x_1 = 0$, $x_2 = 0$, so the estimated mean number of units sold in Region A is

$$\hat{y} = b_0 + b_1(0) + b_2(0) = b_0$$

Observations corresponding to Region B are coded $x_1 = 1$, $x_2 = 0$, so the estimated mean number of units sold in Region B is

$$\hat{y} = b_0 + b_1(1) + b_2(0) = b_0 + b_1$$

Observations corresponding to Region C are coded $x_1 = 0$, $x_2 = 1$, so the estimated mean number of units sold in Region C is

$$\hat{y} = b_0 + b_1(0) + b_2(1) = b_0 + b_2$$

Thus, b_0 is the estimated mean sales for Region A, b_1 is the estimated difference between the mean number of units sold in Region B and the mean number of units sold in Region A, and b_2 is the estimated difference between the mean number of units sold in Region C and the mean number of units sold in Region A.

Two dummy variables were required because sales region is a categorical variable with three levels. But the assignment of $x_1 = 0$ and $x_2 = 0$ to indicate Region A, $x_1 = 1$ and $x_2 = 0$ to indicate Region B, and $x_1 = 0$ and $x_2 = 1$ to indicate Region C was arbitrary. For example, we could have chosen to let $x_1 = 1$ and $x_2 = 0$ indicate Region A, $x_1 = 0$ and $x_2 = 0$ indicate Region B, and $x_1 = 0$ and $x_2 = 1$ indicate Region C. In this case, b_0 is the mean or expected value of sales for Region B, b_1 is the difference between the mean number of units sold in Region A and the mean number of units sold in Region B, and b_2 is the difference between the mean number of units sold in Region C and the mean number of units sold in Region B.

The important point to remember is that when a categorical variable has k levels, $k - 1$ dummy variables are required in the multiple linear regression analysis. Thus, if the sales region example had a fourth region, labeled D, three dummy variables would be necessary. For example, these three dummy variables could then be coded as follows:

$$x_1 = \begin{cases} 1 \text{ if sales Region B} \\ 0 \text{ otherwise} \end{cases} \qquad x_2 = \begin{cases} 1 \text{ if sales Region C} \\ 0 \text{ otherwise} \end{cases} \qquad x_3 = \begin{cases} 1 \text{ if sales Region D} \\ 0 \text{ otherwise} \end{cases}$$

Notes + Comments

1. Detecting multicollinearity when a categorical variable is involved is difficult. The correlation coefficient that we used in Section 8.5 is appropriate only when assessing the relationship between two quantitative variables. However, recall that if any estimated parameters $b_1, b_2, \ldots, b_q$ or associated p values change dramatically when a new independent variable is added to the model (or an existing independent variable is removed from the model), multicollinearity is likely present. We can use our understanding of these ramifications of multicollinearity to assess whether there is multicollinearity that involves a dummy variable. We estimate the linear regression model twice; once with the dummy variable included as an independent variable and once with the dummy variable omitted from the linear regression model. If we see relatively little change in the estimated parameters $b_1, b_2, \ldots, b_q$ or associated p values for the independent variables that have been included in both linear regression models, we can be confident that there is not strong multicollinearity involving the dummy variable.

2. Dummy variables are often used to model seasonal effects in sales data. If the data are collected quarterly and we use winter as the reference season, we may use three dummy variables defined in the following manner:

$$x_1 = \begin{cases} 1 \text{ if spring} \\ 0 \text{ otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ if summer} \\ 0 \text{ otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 \text{ if fall} \\ 0 \text{ otherwise} \end{cases}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8.7 Modeling Nonlinear Relationships

DATA *file*
reynolds

Regression models that are linear combinations of variables may be used to model more complex types of relationships. To illustrate, let us consider the problem facing Reynolds, Inc., a manufacturer of industrial scales and laboratory equipment. Managers at Reynolds want to investigate the relationship between length of employment of their salespeople and the number of electronic laboratory scales sold. The file *reynolds* gives the number of scales sold by 15 randomly selected salespeople for the most recent sales period and the number of months each salesperson has been employed by the firm. Figure 8.25, the scatter chart for these data, indicates a possible curvilinear relationship between the length of time employed and the number of units sold.

Before considering how to develop a curvilinear relationship for Reynolds, let us consider the Excel output in Figure 8.26 for a simple linear regression. For these data, the estimated linear regression equation is

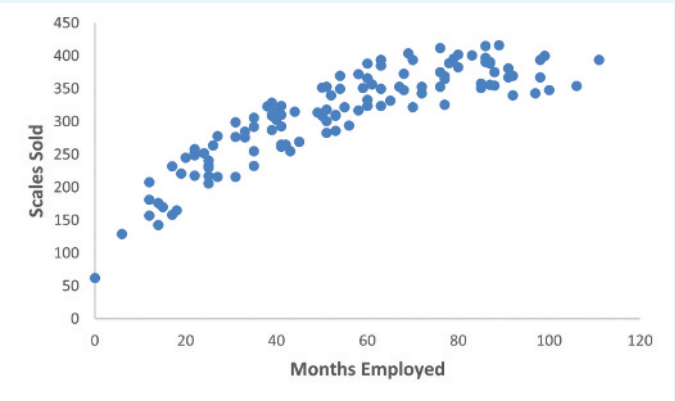
$$\text{Sales} = 183.2914 + 2.3580 \text{ Months Employed}$$

The scatter chart of residuals against the independent variable Months Employed would also suggest that a curvilinear relationship may provide a better fit to the data.

The computer output shows that the relationship is significant $(p \text{ value} = 6.52964\text{E-}39$ in cell E18 of Figure 8.26 for the t test that $\beta_1 = 0)$ and that a linear relationship explains a high percentage of the variability in sales $(r^2 = 0.7564$ in cell B5. However, Figure 8.27 reveals a pattern in the scatter chart of residuals against the predicted values of the dependent variable that suggests that a curvilinear relationship may provide a better fit to the data.

If we have a practical reason to suspect a curvilinear relationship between number of electronic laboratory scales sold by a salesperson and the number of months the salesperson has been employed, we may wish to consider an alternative to simple linear regression. For example, we may believe that a recently hired salesperson faces a learning curve but becomes increasingly more effective over time and that a salesperson who has been in a sales position with Reynolds for a long time eventually becomes burned out and becomes increasingly less effective. If the data supports this theory, Reynolds management can use the model to identify the approximate point in employment when its salespeople begin to lose their effectiveness, and management can plan strategies to counteract salesperson burnout.

Figure 8.25 Scatter Chart for the Reynolds Example

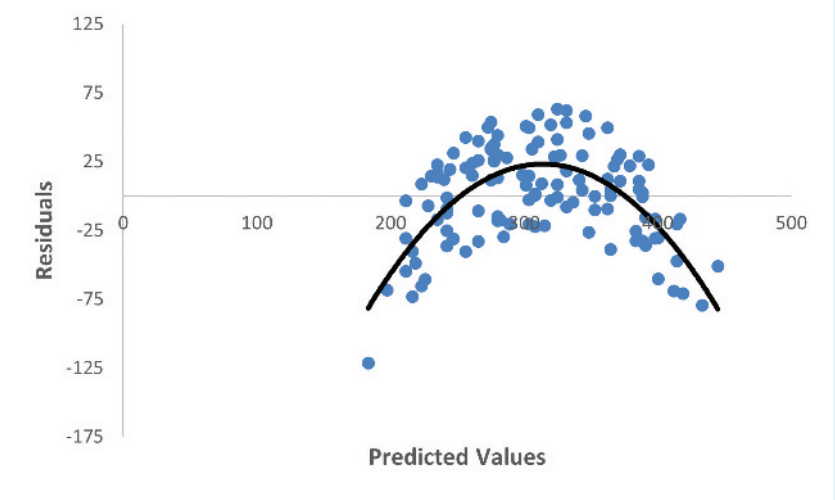


Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.26 Excel Linear Regression Output for the Reynolds Example

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.86970319							
5	R Square	0.75638364							
6	Adjusted R Square	0.75437028							
7	Standard Error	35.2389845							
8	Observations	123							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	466517.3699	466517.3699	375.6825719	6.52964E-39			
13	Residual	121	150256.1098	1241.786031					
14	Total	122	616773.4797						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	183.291425	7.29645984	25.12059677	7.81362E-50	168.8461587	197.7366923	164.1960628	202.3867882
18	Months Employed	2.35795131	0.121653416	19.38253265	6.52964E-39	2.117106288	2.598796338	2.039575531	2.676327095

Figure 8.27 Scatter Chart of the Residuals and Predicted Values of the Dependent Variable for the Reynolds Simple Linear Regression

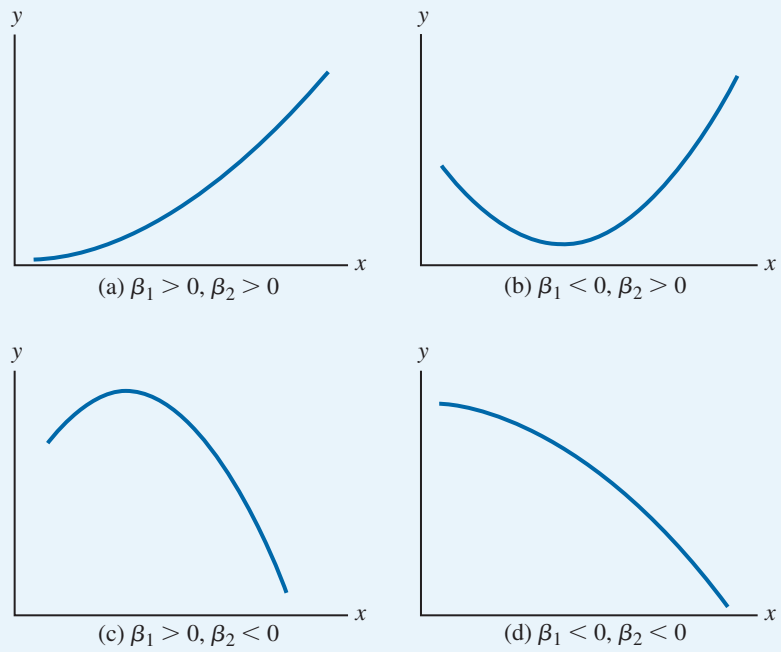


Quadratic Regression Models

To account for the curvilinear relationship between months employed and scales sold that is suggested by the scatter chart of residuals against the predicted values of the dependent variable, we could include the square of the number of months the salesperson

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.28 Relationships That Can Be Fit with a Quadratic Regression Model



If $\beta_2 > 0$, the function is convex (bowl-shaped relative to the x-axis); if $\beta_2 < 0$, the function is concave (mound-shaped relative to the x-axis).

has been employed as a second independent variable in the estimated regression equation:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2 \quad \textbf{(8.18)}$$

Equation (8.18) corresponds to a **quadratic regression model**. As Figure 8.28 illustrates, quadratic regression models are flexible and are capable of representing a wide variety of nonlinear relationships between an independent variable and the dependent variable.

To estimate the values of b_0, b_1, and b_2 in equation (8.18) with Excel, we need to add to the original data the square of the number of months the salesperson has been employed with the firm. Figure 8.29 shows a portion of the Excel spreadsheet that includes the square of the number of months the employee has been with the firm. To create the variable, which we will call MonthsSq, we create a new column and set each cell in that column equal to the square of the associated value of the variable Months. These values are shown in Column B of Figure 8.29.

For the data which Figure 8.29 partially shows, the estimated regression equation corresponding to equation (8.18) is

$$\text{Sales} = 101.9846 + 6.1361\ \text{Months Employed} - 0.0341\ \text{MonthsSq}$$

where MonthsSq is the square of the number of months the salesperson has been employed. Because the value of b_1 (6.1361) is positive, and the value of b_2 (−0.0341) is negative, $\hat{y}$ will initially increase as the number of months the salesperson has been employed increases. As the value of the independent variable Months Employed increases, its squared value increases more rapidly, and eventually $\hat{y}$ will decrease as the number of months the salesperson has been employed increases.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.29 Portion of Data for the Reynolds Quadratic Regression Model

	A	B	C
1	Months Employed	MonthsSq	Scales Sold
2	41	1681	324
3	106	11236	354
4	76	5776	353
5	100	10000	348
6	22	484	249
7	12	144	208
8	85	7225	351
9	111	12321	394
10	40	1600	303
11	51	2601	353
12	0	0	62
13	12	144	157
14	6	36	129
15	56	3136	294
16	19	361	221

The R^2 of 0.8654 indicates that this model explains approximately 86.5% of the variation in Scales Sold for our sample data. The lack of a distinct pattern in the scatter chart of residuals against the predicted values of the dependent variable (Figure 8.31) suggests that the quadratic model fits the data better than the simple linear regression in the Reynolds example. While not shown here, the scatter chart of residuals against the independent variable Months Employed also lack any distinct pattern. We see no marked violations of the conditions necessary for reliable inference, so we will proceed with hypothesis tests of the parameters β_0, β_1, and β_2 for our quadratic regression model.

From the Excel output provided in Figure 8.30, we see that the p values corresponding to the t statistics for Months Employed (1.3091E-30) and MonthsSq (3.7011E-17) are both substantially less than 0.05, and hence we can conclude that the variables Months Employed and MonthsSq are significant. There is a nonlinear relationship between months employed and sales.

Note that if the estimated parameters b_1 and b_2 corresponding to the linear term x and the squared term x^2 are of the same sign, the estimated value of the dependent variable is either increasing over the experimental range of x (when $b_1 > 0$ and $b_2 > 0$) or decreasing over the experimental range of x (when $b_1 < 0$ and $b_2 < 0$). If the estimated parameters b_1 and b_2 corresponding to the linear term x and the squared term x^2 have different signs, the estimated value of the dependent variable has a maximum over the experimental range of x (when $b_1 > 0$ and $b_2 < 0$) or a minimum over the experimental range of x (when $b_1 < 0$ and $b_2 > 0$). In these instances, we can find the estimated maximum or minimum over the experimental range of x by finding the value of x at which the estimated value of the dependent variable stops increasing and begins decreasing (when a maximum exists) or stops decreasing and begins increasing (when a minimum exists). For example, we estimate that when months employed increases by 1 from some value x $(x + 1)$, sales changes by

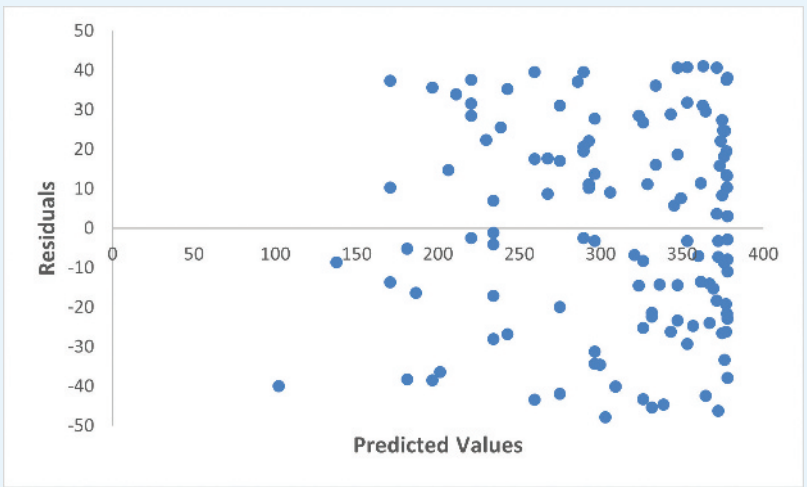
$$\begin{aligned} &6.1361[(x+1) - x] - 0.0341[(x+1)^2 - x^2] \\ &\quad = 6.1361(x - x + 1) - 0.0341(x^2 + 2x + 1 - x^2) \\ &\quad = 6.1361 - 0.0341(2x + 1) \\ &\quad = 6.1020 - 0.0682x \end{aligned}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.30 Excel Output for the Reynolds Quadratic Regression Model

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.930279093							
5	R Square	0.865419191							
6	Adjusted R Square	0.863176178							
7	Standard Error	26.30048695							
8	Observations	123							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	533767.606	266883.803	385.8287969	5.48262E-53			
13	Residual	120	83005.87365	691.7156137					
14	Total	122	616773.4797						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	101.9845664	9.881914965	10.32032422	2.93161E-18	82.41906246	121.5500703	76.11943318	127.8496995
18	Months Employed	6.136102358	0.393784541	15.5823851	1.30911E-30	5.356436373	6.915768344	5.105402375	7.166802342
19	MonthsSq	-0.034104142	0.003458789	-9.860140815	3.70105E-17	-0.040952303	-0.027255981	-0.043157248	-0.025051036

Figure 8.31 Scatter Chart of the Residuals and Predicted Values of the Dependent Variable for the Reynolds Quadratic Regression Model



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

That is, estimated Sales initially increases as Months Employed increases and then eventually decreases as Months Employed increases. Solving this result for x:

$$\begin{aligned} 6.1020 - 0.0682x &= 0 \\ -0.0682x &= -6.1020 \\ x &= \frac{-6.1020}{-0.0682} = 89.5 \end{aligned}$$

tells us that estimated maximum sales occurs at approximately 90 months (in about seven years and six months). We can then find the estimated maximum value of the dependent variable Sales by substituting this value of x into the estimated regression equation:

Polynomial regression models of higher than second or third order are rarely used in business analytics.

$$\text{Sales} = 101.9846 + 6.1361(89.5) - 0.0341(89.5^2) = 377.98$$

At approximately 90 months, the maximum estimated sales of approximately 378 scales occurs.

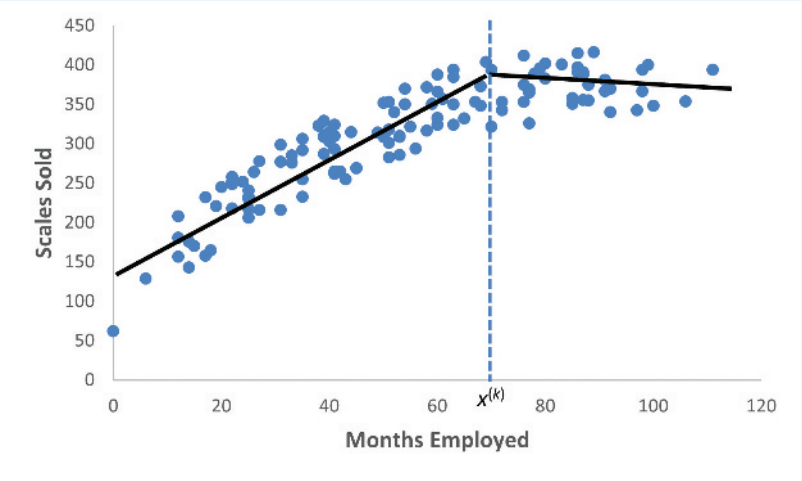
Piecewise Linear Regression Models

As an alternative to a quadratic regression model, we can recognize that below some value of Months Employed, the relationship between Months Employed and Sales appears to be positive and linear, whereas the relationship between Months Employed and Sales appears to be negative and linear for the remaining observations. A **piecewise linear regression model** will allow us to fit these relationships as two linear regression models that are joined at the value of Months at which the relationship between Months Employed and Sales changes.

A piecewise linear regression model is sometimes referred to as a segment regression or a spline model.

Our first step in fitting a piecewise linear regression model is to identify the value of the independent variable Months Employed at which the relationship between Months Employed and Sales changes; this point is called the **knot**, or *breakpoint*. Although theory should determine this value, analysts often use the sample data to aid in the identification of this point. Figure 8.32 provides the scatter chart for the Reynolds data with an indication

Figure 8.32 Possible Position of Knot $x^{(k)} = 70$



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

of the possible location of the knot, which we have denoted $x^{(k)}$. From this scatter chart, it appears that the knot is at approximately 70 months.

Once we have decided on the location of the knot, we define a dummy variable that is equal to zero for any observation for which the value of Months Employed is less than or equal to the value of the knot, and equal to one for any observation for which the value of Months Employed is greater than the value of the knot:

$$x_k = \begin{cases} 0 \text{ if } x_1 \leq x^{(k)} \\ 1 \text{ if } x_1 > x^{(k)} \end{cases} \tag{8.19}$$

where

x_1 = Months
$x^{(k)}$ = the value of the knot (70 months for the Reynolds example)
x_k = the knot dummy variable

We then fit the following estimated regression equation:

$$\hat{y} = b_0 + b_1 x_1 + b_2 (x_1 - x^{(k)}) x_k \tag{8.20}$$

In the Figure 8.34, the variable Knot Dummy(Months – 70) is the product of the corresponding values of knot dummy variable and the difference between Months Employed and the knot value. That is,* C2 = A2*(B2 − 70) *in the spreadsheet.*

A portion of the data is provided in Figure 8.33. The corresponding Excel output for the Reynolds piecewise linear regression model is provided in Figure 8.34. Because we placed the knot at $x^{(k)} = 70$, the estimated regression equation is

$$\hat{y} = 147.4622 + 3.3539 x_1 - 3.6605 (x_1 - 70) x_k$$

The output shows that the p value corresponding to the t statistic for the knot term ($p = 0.0014$) is less than 0.05, and hence we can conclude that adding the knot to the model with Months Employed as the independent variable is significant.

But what does this model mean? For any value of Months less than or equal to 70, the knot term $3.6605(x_1 - 70)x_k$ is zero because the knot dummy variable $x_k = 0$, so the regression equation is

$$\hat{y} = 147.4622 + 3.3539 x_1$$

For any value of Months Employed greater than 70, the knot term is $-3.6605(x_1 - 70)$ because the knot dummy variable $x_k = 1$, so the regression equation is

$$\begin{aligned} \hat{y} &= 147.4622 + 3.3539 x_1 - 3.6605 (x_1 - 70) \\ &= 147.4622 - 3.6605(-70) + (3.3539 - 3.6605) x_1 = 403.6995 - 0.3066 x_1 \end{aligned}$$

Note that if Months Employed is equal to 70, both linear regression models yield the same value of $\hat{y}$:

$$\hat{y} = 147.4622 + 3.3539(70) = 403.6995 - 0.3066(70) = 382.2374$$

Multiple knots can be used to fit complex piecewise linear regression models.

So the two regression segments are joined at the knot.

The interpretation of this model is similar to the interpretation of the quadratic regression model. A salesperson's sales are expected to increase by 3.4094 electronic laboratory scales for each month of employment until the salesperson has been employed for 70 months. At that point the salesperson's sales are expected to decrease by 4.4632 electronic laboratory scales for each additional month of employment.

Should we use the quadratic regression model or the piecewise linear regression model? These models fit the data equally well, and both have reasonable interpretations, so we cannot differentiate between the models on either of these criteria. Thus, we must consider whether the abrupt change in the relationship between Sales and Months Employed that is suggested by the piecewise linear regression model captures the real relationship between Sales and Months Employed better than the smooth change in the relationship between Sales and Months Employed suggested by the quadratic model.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.33 Data and Excel Output for the Reynolds Piecewise Linear Regression Model

	A	B	C	D
1	Knot	Months Employed	Knot Dummy *(Months-70)	Scales Sold
2	0	41	0	324
3	1	106	36	354
4	1	76	6	353
5	1	100	30	348
6	0	22	0	249
7	0	12	0	208
8	1	85	15	351
9	1	111	41	394
10	0	40	0	303
11	0	51	0	353
12	0	0	0	62
13	0	12	0	157
14	0	6	0	129
15	0	56	0	294
16	0	19	0	221

Interaction Between Independent Variables

Often the relationship between the dependent variable and one independent variable is different at various values of a second independent variable. When this occurs, it is called an **interaction**. If the original data set consists of observations for y and two independent variables x_1 and x_2, we can incorporate an x_1x_2 interaction into the estimated multiple linear regression equation in the following manner:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 \tag{8.21}$$

To provide an illustration of an interaction and what it means, let us consider the regression study conducted by Tyler Personal Care for one of its new shampoo products. The two factors believed to have the most influence on sales are unit selling price and advertising expenditure. To investigate the effects of these two variables on sales, prices of $2.00, $2.50, and $3.00 were paired with advertising expenditures of $50,000 and $100,000 in 24 test markets.

DATA *file*
tyler

The data collected by Tyler are provided in the file *tyler*. Figure 8.35 shows the sample mean sales for the six price and advertising expenditure combinations. Note that the sample

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.34 Excel Output for the Reynolds Piecewise Linear Regression Model

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.918413194							
5	R Square	0.843482795							
6	Adjusted R Square	0.840874175							
7	Standard Error	28.36307185							
8	Observations	123							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	520237.8183	260118.9092	323.3444374	4.71764E-49			
13	Residual	120	96535.66134	804.4638445					
14	Total	122	616773.4797						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	147.4621856	7.328929534	20.12056262	2.92532E-40	132.9514152	161.972956	128.2792905	166.6450808
18	Months Employed	3.353930564	0.15634062	21.45271378	7.01772E-43	3.044387018	3.66347411	2.944721321	3.763139807
19	Knot Dummy *(Months-70)	-3.660533511	0.447948243	-8.171777807	3.54771E-13	-4.547439858	-2.773627165	-4.833002715	-2.488064308

In the file tyler, *the independent variable Price is in column A, the independent variable Advertising Expenditures is in column B, and the dependent variable Sales is in column D. We created the interaction variable Price*Advertising in column C by entering the function =A2*B2 in cell C2, and then copying cell C2 into cells C3 through C25.*

mean sales corresponding to a price of $2.00 and an advertising expenditure of $50,000 is 461,000 units and that the sample mean sales corresponding to a price of $2.00 and an advertising expenditure of $100,000 is 808,000 units. Hence, with price held constant at $2.00, the difference in mean sales between advertising expenditures of $50,000 and $100,000 is $808,000 - 461,000 = 347,000$ units. When the price of the product is $2.50, the difference in mean sales between advertising expenditures of $50,000 and $100,000 is $646,000 - 364,000 = 282,000$ units. Finally, when the price is $3.00, the difference in mean sales between advertising expenditures of $50,000 and $100,000 is $375,000 - 332,000 = 43,000$ units. Clearly, the difference between mean sales for advertising expenditures of $50,000 and mean sales for advertising expenditures of $100,000 depends on the price of the product. In other words, at higher selling prices, the effect of increased advertising expenditure diminishes. These observations provide evidence of interaction between the price and advertising expenditure variables.

When interaction between two variables is present, we cannot study the relationship between one independent variable and the dependent variable y independently of the other variable. In other words, meaningful conclusions can be developed only if we consider the joint relationship that both independent variables have with the dependent variable. To account for the interaction, we use equation (8.21), where

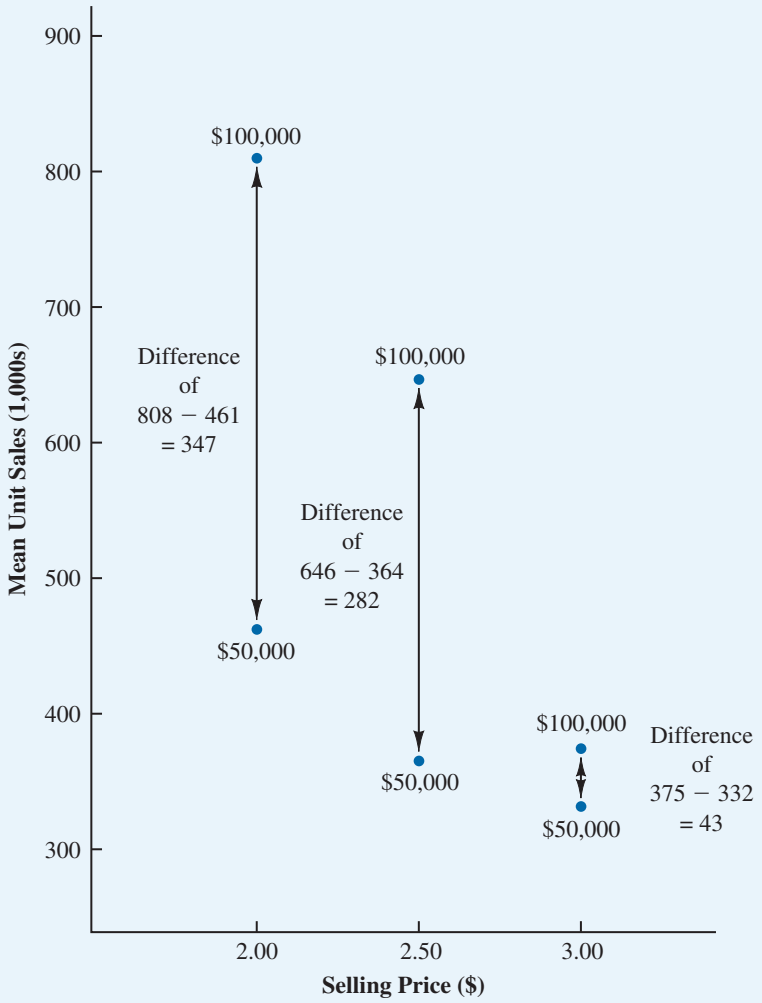
$$y = \text{Unit Sales (1000s)}$$
$$x_1 = \text{Price (\$)}$$
$$x_2 = \text{Adverstising Expenditure (\$1000s)}$$

Note that equation (8.21) reflects Tyler's belief that the number of units sold is related to selling price and advertising expenditure (accounted for by the $\beta_1 x_1$ and $\beta_2 x_2$ terms) and an interaction between the two variables (accounted for by the $\beta_3 x_1 x_2$ term).

The Excel output corresponding to the interaction model for the Tyler Personal Care example is provided in Figure 8.36.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.35 Mean Unit Sales (1,000s) as a Function of Selling Price and Advertising Expenditures



The resulting estimated regression equation is

$$\text{Sales} = -275.8333 + 175\ \text{Price} + 19.68\ \text{Advertising} - 6.08\ \text{Price*Advertising}$$

Because the *p* value corresponding to the *t* test for Price*Advertising is close to zero (8.6772E-10), we conclude that interaction is significant. Thus, the relationship between advertising expenditure and sales depends on the price (and the relationship between price and sales depends on advertising expenditure).

Our initial review of these results may alarm us: How can price have a positive estimated regression coefficient? With the exception of luxury goods, we expect sales to decrease as price increases. Although this result appears counterintuitive, we can make sense of this model if we work through the interpretation of the interaction. In other words, the relationship between the independent variable Price and the dependent variable Sales is different at various values of Advertising (and the relationship between the independent variable Advertising and the dependent variable Sales is different at various values of Price).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.36 Excel Output for the Tyler Personal Care Linear Regression Model with Interaction

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.988993815							
5	R Square	0.978108766							
6	Adjusted R Square	0.974825081							
7	Standard Error	28.17386496							
8	Observations	24							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	709316	236438.6667	297.8692311	9.25881E-17			
13	Residual	20	15875.33333	793.7666667					
14	Total	23	725191.3333						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	-275.8333333	112.8421033	-2.444418575	0.023898351	-511.2178361	-40.44883053	-596.9074508	45.24078413
18	Price	175	44.54679188	3.928453489	0.0008316	82.07702045	267.9229796	48.24924412	301.7507559
19	Advertising Expenditure ($1000s)	19.68	1.42735225	13.78776683	1.1263E-11	16.70259538	22.65740462	15.61869796	23.74130204
20	Price*Advertising	-6.08	0.563477299	-10.79014187	8.67721E-10	-7.255393049	-4.904606951	-7.683284335	-4.476715665

It becomes easier to see how the predicted value of Sales depends on Price by using the estimated regression equation to consider the effect when Price increases by $1:

$$\text{Sales After \$1 Price Increase} = -275.8333 + 175\,(\text{Price} + 1) + 19.68\,\text{Advertising} - 6.08\,(\text{Price} + 1) * \text{Advertising}$$

Thus,

$$\text{Sales After \$1 Price Increase} - \text{Sales Before \$1 Price Increase} = 175 - 6.08 * \text{Advertising Expenditure}$$

So the change in the predicted value of sales when the independent variable Price increases by $1 depends on how much was spent on advertising.

Consider a concrete example. If Advertising Expenditures is $50,000 when price is $2.00, we estimate sales to be

$$\text{Sales} = -275.8333 + 175(2) + 19.68(50) - 6.08(2)(50) = 450.1667, \text{or } 450{,}167 \text{ units}$$

At the same level of Advertising Expenditures ($50,000) when price is $3.00, we estimate sales to be

$$\text{Sales} = -275.8333 + 175(3) + 19.68(50) - 6.08(3)(50) = 321.1667, \text{or } 321{,}167 \text{ units}$$

So when Advertising Expenditures is $50,000, a change in price from $2.00 to $3.00 results in a $450{,}167 - 321{,}167 = 129{,}000$ unit decrease in estimated sales. However, if Advertising Expenditures is $100,000 when price is $2.00, we estimate sales to be

$$\text{Sales} = -275.8333 + 175(2) + 19.68(100) - 6.08(2)(100) = 826.1667, \text{or } 826{,}167 \text{ units}$$

At the same level of Advertising Expenditures ($100,000) when price is $3.00, we estimate sales to be

$$\text{Sales} = -275.8333 + 175(3) + 19.68(100) - 6.08(3)(100) = 393.1667, \text{or } 393{,}167 \text{ units}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

So when Advertising Expenditures is \$100,000, a change in price from \$2.00 to \$3.00 results in a $826{,}167 - 393{,}167 = 433{,}000$ unit decrease in estimated sales. When Tyler spends more on advertising, its sales are more sensitive to changes in price. Perhaps at larger Advertising Expenditures, Tyler attracts new customers who have been buying the product from another company and so are more aware of the prices charged for the product by Tyler's competitors.

There is a second and equally valid interpretation of the interaction; it tells us that the relationship between the independent variable Advertising Expenditure and the dependent variable Sales is different at various values of Price. Using the estimated regression equation to consider the effect when Advertising Expenditure increases by \$1,000:

$$\begin{aligned}\text{Sales After \$1K Advertising Increase} = {} & -275.8333 + 175\,\text{Price} + 19.68\,(\text{Advertising} + 1) \\ & -6.08\,\text{Price}^{*}\,(\text{Advertising} + 1)\end{aligned}$$

Thus,

$$\begin{aligned}\text{Sales After \$1K Advertising Increase} - \text{Sales Before \$1K Advertising Increase} = {} & 19.68 \\ & - 6.08\,\text{Price}\end{aligned}$$

So the change in the predicted value of the dependent variable that occurs when the independent variable Advertising Expenditure increases by \$1,000 depends on the price.

Thus, if Price is \$2.00 when Advertising Expenditures is \$50,000, we estimate sales to be

$$\text{Sales} = -275.8333 + 175(2) + 19.68(50) - 6.08(2)(50) = 450.1667, \text{or } 450{,}167 \text{ units}$$

At the same level of Price (\$2.00) when Advertising Expenditures is \$100,000, we estimate sales to be

$$\text{Sales} = -275.8333 + 175(2) + 19.68(100) - 6.08(2)(100) = 826.1667, \text{or } 826{,}167 \text{ units}$$

So when Price is \$2.00, a change in Advertising Expenditures from \$50,000 to \$100,000 results in a $826{,}167 - 450{,}167 = 376{,}000$ unit increase in estimated sales. However, if Price is \$3.00 when Advertising Expenditures is 50,000, we estimate sales to be

$$\text{Sales} = -275.8333 + 175(3) + 19.68(50) - 6.08(3)(50) = 321.1667, \text{ or } 321{,}167 \text{ units}$$

At the same level of Price (\$3.00) when Advertising Expenditures is \$100,000, we estimate sales to be

$$\text{Sales} = -275.8333 + 175(3) + 19.68(100) - 6.08(3)(100) = 393.1667, \text{ or } 393{,}167 \text{ units}$$

So when Price is \$3.00, a change in Advertising Expenditure from \$50,000 to \$100,000 results in a $393.167 - 321{,}167 = 72{,}000$ unit increase in estimated sales. When the price of Tyler's product is high, its sales are less sensitive to changes in advertising expenditure. Perhaps as Tyler increases its price, it must advertise more to convince potential customers that its product is a good value.

Notes + Comments

1. Just as a dummy variable can be used to allow for different *y*-intercepts for the two groups represented by the dummy, we can use an interaction between a dummy variable and a quantitative independent variable to allow for different relationships between independent and dependent variables for the two groups represented by the dummy. Consider the Butler Trucking example: Travel time is the dependent variable *y*, miles traveled and number of deliveries are the quantitative independent variables x_1 and x_2, and the dummy variable x_3 differentiates between driving assignments that included travel on a congested segment of a highway and driving assignments that did not. If we believe that the relationship between miles traveled and travel time differs for driving assignments that included travel on a congested segment of a highway and those that did not, we could create a new variable that is the interaction between miles traveled and the dummy variable $(x_4 = x_1^{*}x_3)$ and estimate the following model:

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_4$$

If a driving assignment does not include travel on a congested segment of a highway, $x_4 = x_1 * x_3 = x_1 * 0 = 0$ and the model is

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

If a driving assignment does include travel on a congested segment of a highway, $x_4 = x_1 * x_3 = x_1 * 1 = x_1$ and the model is

$$\begin{aligned}\hat{y} &= b_0 + b_1x_1 + b_2x_2 + b_3x_1(1)\\ &= b_0 + (b_1 + b_3)x_1 + b_2x_2\end{aligned}$$

So in this model b_1 is the estimate of the relationship between miles traveled and travel time for driving assignments that do not include travel on a congested segment of a highway, and $b_1 + b_3$ is the estimate of the relationship between miles traveled and travel time for driving assignments that do include travel on a congested segment of a highway.

2. Multicollinearity can be divided into two types. *Data-based multicollinearity* occurs when separate independent variables that are related are included in the model, whereas *structural multicollinearity* occurs when a new independent variable is created by taking a function of one or more existing independent variables. If we use ratings that consumers give on bread's aroma and taste as independent variables in a model for which the dependent variable is the overall rating of the bread, the multicollinearity that would exist between the aroma and taste ratings is an example of data-based multicollinearity. If we build a quadratic model for which the independent variables are ratings that consumers give on bread's aroma and the square of the ratings that consumers give on bread's aroma, the multicollinearity that would exist is an example of structural multicollinearity.
3. Structural multicollinearity occurs naturally in polynomial regression models and regression models with interactions. You can greatly reduce the structural multicollinearity in a polynomial regression by centering the independent variable x (using $x - \bar{x}$ in place of x). In a regression model with interaction, you can greatly reduce the structural multicollinearity by centering both independent variables that interact. However, quadratic regression models and regression models with interactions are frequently used only for prediction; in these instances centering independent variables is not necessary because we are not concerned with inference.
4. Note that we can combine a quadratic effect with interaction to produce a second-order polynomial model with interaction between the two independent variables. The resulting estimated regression equation is

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_2^2 + b_5x_1x_2$$

This model provides a great deal of flexibility in capturing nonlinear effects.

5. The parameters in a quadratic regression model, a piecewise linear regression model, and a regression model with interactions can all be estimated using the same least squares method because while the variables may represent nonlinear functions, the models are still linear with respect to the parameters.

8.8 Model Fitting

Finding an effective linear regression model can be challenging. Although we rely on theory to guide us, often we are faced with a large number of potential independent variables from which to choose. In this section, we discuss common methods for building a linear regression model and the potential hazards of these approaches.

Variable Selection Procedures

When there are many independent variables to consider, special procedures are sometimes employed to select the independent variables to include in the linear regression model. These variable selection procedures include **backward elimination**, **forward selection**, **stepwise selection**, and the **best subsets** procedure. Given a data set with several possible independent variables, we can use these procedures to identify which independent variables provide a model that best satisfies some criterion. The first three procedures are iterative; at each step of the procedure a single independent variable is added or removed and the new model is evaluated. The process continues until a stopping criterion indicates that the procedure cannot find a superior model. The best subsets procedure is not a one-variable-at-a-time procedure; it evaluates linear regression models involving different subsets of the independent variables.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The backward elimination procedure begins with the linear regression model that includes all of the independent variables under consideration. At each step of the procedure, backward elimination considers the removal of an independent variable according to some criterion. One such criterion is to check if any independent variables currently in the model are not significant at a specified level of significance, and if so, then remove the least significant of these independent variables from the model. The linear regression model is then refit with the remaining independent variables and statistical significance is reexamined. The backward elimination procedure stops when all independent variables in the model are significant at a specified level of significance.

The forward selection procedure begins with none of the independent variables under consideration included in the linear regression model. At each step of the procedure, forward selection considers the addition of an independent variable according to some criterion. One such criterion is to check if any independent variables currently not in the model would be significant at a specified level of significance if included, and if so, then add the most significant of these independent variables to the model. The linear regression model is then refit with the additional independent variable and statistical significance is reexamined. The forward selection procedure stops when all of the independent variables not in the model would not be significant at a specified level of significance if included in the model.

The stepwise procedure requires that the criterion for an independent variable to enter the linear regression model is more difficult to satisfy than the criterion for an independent variable to be removed from the linear regression model. This requirement prevents the same independent variable from exiting and then reentering the linear regression model in the same step.

Similar to the forward selection procedure, the stepwise procedure begins with none of the independent variables under consideration included in the linear regression model. The analyst establishes both a criterion for allowing independent variables to enter the model and a criterion for allowing independent variables to remain in the model. One such criterion adds the most significant variable and removes the least significant variable at each iteration. To initiate the procedure, the most significant independent variable is added to the empty model if its level of significance satisfies the entering threshold. Each subsequent step involves two intermediate steps. First, the remaining independent variables not in the current model are evaluated, and the most significant one is added to the model if its significance satisfies the threshold to remain in the model. Then the independent variables in the resulting model are evaluated, and the least significant variable is removed if its level of significance fails to satisfy the threshold to remain in the model. The procedure stops when no independent variable not currently in the model has a level of significance that satisfies the entering threshold, and no independent variable currently in the model has a level of significance that fails to satisfy the threshold to remain in the model.

In the best subsets procedure, simple linear regression models for each of the independent variables under consideration are generated, and then the multiple linear regression models with all combinations of two independent variables under consideration are generated, and so on. Once a linear regression model has been generated for every possible subset of the independent variables under consideration, the entire collection of linear regression models can be compared and evaluated by the analyst.

Although these algorithms are potentially useful when dealing with a large number of potential independent variables, they do not necessarily provide useful models. Once the procedure terminates, you should deliberate whether the combination of independent variables included in the final linear regression model makes sense from a practical standpoint and consider whether you can create a more useful linear regression model with more meaningful interpretation through the addition or removal of independent variables. Use your own judgment and intuition about your data to refine the results of these algorithms.

Overfitting

The objective in building a linear regression model (or any other type of mathematical model) is to provide the simplest accurate representation of the population. A model that is relatively simple will be easy to understand, interpret, and use, and a model that accurately represents the population will yield meaningful results.

When we base a model on sample data, we must be wary. Sample data generally do not perfectly represent the population from which they are drawn; if we attempt to fit a model

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

too closely to the sample data, we risk capturing behavior that is idiosyncratic to the sample data rather than representative of the population. When the model is too closely fit to sample data and as a result does not accurately reflect the population, the model is said to have been overfit.

The principle of using the simplest meaningful model possible without sacrificing accuracy is referred to as Ockham's razor, the law of parsimony, or the law of economy.

Overfitting generally results from creating an overly complex model to explain idiosyncrasies in the sample data. In regression analysis, this often results from the use of complex functional forms or independent variables that do not have meaningful relationships with the dependent variable. If a model is overfit to the sample data, it will perform better on the sample data used to fit the model than it will on other data from the population. Thus, an overfit model can be misleading with regard to its predictive capability and its interpretation.

Overfitting is a difficult problem to detect and avoid, but there are strategies that can help mitigate this problem. Use only independent variables that you expect to have real and meaningful relationships with the dependent variable. Use complex models, such as quadratic regression models and piecewise linear regression models, only when you have a reasonable expectation that such complexity provides a more accurate depiction of what you are modeling. Do not let software dictate your model. Use iterative modeling procedures, such as the stepwise and best-subsets procedures, only for guidance and not to generate your final model. Use your own judgment and intuition about your data and what you are modeling to refine your model. If you have access to a sufficient quantity of data, assess your model on data other than the sample data that were used to generate the model (this is referred to as **cross-validation**).

One simple cross-validation approach is the **holdout method**. In the holdout method, the sample data are randomly divided into mutually exclusive and collectively exhaustive training and validation sets. The **training set** is the data set used to build the candidate models that appear to make practical sense. The **validation set** is the set of data used to compare model performances and ultimately select a model for predicting values of the dependent variable. For example, we might randomly select half of the data for use in developing linear regression models. We could use these data as our training set to estimate a model or a collection of models that appear to perform well. Then we use the remaining half of the data as a validation set to assess and compare the models' performances and ultimately select the model that minimizes some measure of overall error when applied to the validation set. The advantages of the holdout method are that it is simple and quick. However, results of a holdout sample can vary greatly depending on which observations are randomly selected for the training set, the number of observations in the sample, and the number of observations that are randomly selected for the training and validation sets.

We provide a more in-depth discussion of holdout methods in Chapter 10.

8.9 Big Data and Linear Regression

Inference and Very Large Samples

DATA*file*
largecredit

Consider the example of a credit card company that has a very large database of information provided by its customers when they apply for credit cards. These customer records include information on the customer's annual household income, number of years of post–high school education, and number of members of the customer's household. In a second database, the company has records of the credit card charges accrued by each customer over the past year. Because the company is interested in using annual household income, the number of years of post–high school education, and the number of members of the household reported by new applicants to predict the credit card charges that will be accrued by these applicants, a data analyst links these two databases to create one data set containing all relevant information for a sample of 5,000 customers. The file *largecredit* contains these data, split into a training set of 3,000 observations and a validation set of 2,000 observations.

The company has decided to apply multiple linear regression to these data to develop a model for predicting annual credit card charges for its new applicants. The dependent variable in the model is credit card charges accrued by a customer in the data set over the past

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

year (y); the independent variables are the customer's annual household income (x_1), number of members of the household (x_2), and number of years of post–high school education (x_3). Figure 8.37 provides Excel output for the multiple linear regression model based on the 3,000 observations in the training set.

The model has a coefficient of determination of 0.3632 (see cell B5 in Figure 8.37), indicating that this model explains approximately 36% of the variation in credit card charges accrued by the customers in the sample over the past year. The p value for each test of the individual parameters is also very small (see cells E18 through E20), indicating that for each independent variable we can reject the hypothesis of no relationship with the dependent variable. The estimated slopes associated with the dependent variables are all highly significant. The model estimates the following:

- For a fixed number of household members and number of years of post–high school education, accrued credit card charges increase by $121.34 when a customer's annual household income increases by $1,000. This is shown in cell B18 of Figure 8.37.
- For a fixed annual household income and number of years of post–high school education, accrued credit card charges increase by $528.10 when a customer's household increases by one member. This is shown in cell B19 of Figure 8.37.
- For a fixed annual household income and number of household members, accrued credit card charges decrease by $535.36 when a customer's number of years of post–high school education increases by one year. This is shown in cell B20 of Figure 8.37.

Because the y-intercept is an obvious result of extrapolation (no customer in the data has values of zero for annual household income, number of household members, *and* number of years of post–high school education), the estimated parameter β_0 is meaningless.

The small p values associated with a model that is fit on an extremely large sample do not imply that an extremely large sample solves all problems. Virtually all relationships between independent variables and the dependent variable will be statistically significant if the sample size is sufficiently large. That is, if the sample size is very large, there will be little difference in the b_j values generated by different random samples. Because we address the variability in potential values of our estimators through the use of statistical inference, and variability of our estimates b_j essentially disappears as the sample size grows very large, inference is of little use for estimates generated from very large samples. Thus, we

Figure 8.37 Excel Regression Output for Credit Card Company Example

	A	B	C	D	E	F	G	H	I
3	*Regression Statistics*								
4	Multiple R	0.602663145							
5	R Square	0.363202867							
6	Adjusted R Square	0.362565219							
7	Standard Error	4834.449957							
8	Observations	3000							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	39937797910	13312599303	569.5983495	6.5207E-293			
13	Residual	2996	70022231537	23371906.39					
14	Total	2999	1.0996E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	2119.600282	333.0922952	6.363402314	2.27497E-10	1466.487528	2772.713036	1261.064442	2978.136122
18	Annual Income ($1000)	121.3384676	3.165148859	38.33578544	5.4905E-262	115.1323826	127.5445525	113.1803871	129.496548
19	Household Size	528.0996852	42.84154037	12.32681366	4.29401E-34	444.097873	612.1014973	417.6768433	638.522527
20	Years of Post-High School Education	-535.3593516	58.5960221	-9.136445316	1.15792E-19	-650.2518601	-420.4668432	-686.3889184	-384.3297849

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The phenomenon by which the value of an estimate generally becomes closer to the value of the parameter being estimated as the sample size grows is called the Law of Large Numbers.

generally are not concerned with the conditions a linear regression model must satisfy in order for inference to be reliable when we use a very large sample. Multicollinearity, on the other hand, can result in confusing or misleading parameters $b_1, b_2, \ldots, b_q$ and so is still a concern when we use a large data set to estimate a linear regression model that is to be used for explanatory purposes.

How much does sample size matter? Table 8.4 provides the parameter estimates and the corresponding p values for multiple linear regression models estimated on the first 50 observations, the second 50 observations, and so on for the *largecredit* data. Note that, even though the means of the parameter estimates for the linear regression models based on 50 observations are similar to the parameter estimates based on the full sample of 5,000 observations, the individual values of the estimated parameters in the linear regression models based on 50 observations show a great deal of variation. In these 10 linear regression models, the estimated values of b_0 range from −2,191.590 to 8,994.040, the estimated values of b_1 range from 73.207 to 155.187, the estimated values of b_2 range from −489.932 to 1,267.041, and the estimated values of b_3 range from −974.791 to 207.828. This is reflected in the p values corresponding to the parameter estimates in the linear regression models based on 50 observations, which are substantially larger than the corresponding p values in the linear regression model based on 3,000 observations. These results underscore the impact that a very large sample size can have on inference.

For another example, suppose the credit card company also has a separate database of information on shopping and lifestyle characteristics that it has collected from its customers during a recent Internet survey. The data analyst notes in the results in Figure 8.37 that the original linear regression model fails to explain almost 65% of the variation in credit card charges accrued by the customers in the data set. In an attempt to increase the variation in the dependent variable explained by the model, the data analyst decides to augment the original linear regression model with a new independent variable, number of hours per week spent watching television (which we will designate as x_4). The analyst runs the new multiple linear regression analysis and achieves the results shown in Figure 8.38.

The new model has a coefficient of determination of 0.3645 (see cell B5 in Figure 8.38), indicating the addition of number of hours per week spent watching television increased the explained variation in sample values of accrued credit card charges by less than 1%. The estimated parameters and associated p values for annual household income, number of household members, and number of years of post–high school education changed little after introducing into the model the number of hours per week spent watching television.

Table 8.4 Parameter Estimates and the Corresponding p Values for 10 Multiple Linear Regression Models, Each Estimated on 50 Observations from the *LargeCredit* Data

Observations	b_0	p Value	b_1	p Value	b_2	p Value	b_3	p Value
1–50	−805.152	0.7814	154.488	1.45E-06	234.664	0.5489	207.828	0.6721
51–100	894.407	0.6796	125.343	2.23E-07	822.675	0.0070	−355.585	0.3553
101–150	−2,191.590	0.4869	155.187	3.56E-07	674.961	0.0501	−25.309	0.9560
151–200	2,294.023	0.3445	114.734	1.26E-04	297.011	0.3700	−537.063	0.2205
201–250	8,994.040	0.0289	103.378	6.89E-04	−489.932	0.2270	−375.601	0.5261
251–300	7,265.471	0.0234	73.207	1.02E-02	−77.874	0.8409	−405.195	0.4060
301–350	2,147.906	0.5236	117.500	1.88E-04	390.447	0.3053	−374.799	0.4696
351–400	−504.532	0.8380	118.926	8.54E-07	798.499	0.0112	45.259	0.9209
401–450	1,587.067	0.5123	81.532	5.06E-04	1,267.041	0.0004	−891.118	0.0359
451–500	−315.945	0.9048	148.860	1.07E-05	1,000.243	0.0053	−974.791	0.0420
Mean	1,936.567		119.316		491.773		−368.637	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.38 Excel Regression Output for Credit Card Company Example After Adding Number of Hours per Week Spent Watching Television

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.603724482							
5	R Square	0.36448325							
6	Adjusted R Square	0.36363448							
7	Standard Error	4830.393498							
8	Observations	3000							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	4	40078588918	10019647230	429.4250838	8.3277E-293			
13	Residual	2995	69881440529	23332701.35					
14	Total	2999	1.0996E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	1712.552073	371.7837807	4.606311953	4.26973E-06	983.5746542	2441.529492	754.2898349	2670.814311
18	Annual Income ($1000)	121.6120724	3.164453912	38.43066631	4.943E-263	115.4073492	127.8167955	113.4557814	129.7683633
19	Household Size	531.213362	42.82435656	12.40446803	1.71315E-34	447.2452317	615.1814922	420.8347874	641.5919365
20	Years of Post-High School Education	-539.8345703	58.57519443	-9.216095235	5.64208E-20	-654.6862563	-424.9828843	-690.8104864	-388.8586541
21	Hours Per Week Watching Television	12.55178379	5.109759992	2.456433142	0.014088759	2.532789303	22.57077828	-0.618478873	25.72204645

The estimated parameter for number of hours per week spent watching television is 12.55 (see cell B21 in Figure 8.38), suggesting that a 1-hour increase coincides with an increase of $12.55 in credit card charges accrued by each customer over the past year. The *p* value associated with this estimate is 0.014 (see cell E21 in Figure 8.38), so, at a 5% level of significance, we can reject the hypothesis that there is no relationship between the number of hours per week spent watching television and credit card charges accrued. However, when the model is based on a very large sample, almost all relationships will be significant whether they are real or not, and statistical significance does not necessarily imply that a relationship is meaningful or useful.

Is it reasonable to expect that the credit card charges accrued by a customer are related to the number of hours per week the consumer watches television? If not, the model that includes number of hours per week the consumer watches television as an independent variable may provide inaccurate or unreliable predictions of the credit card charges that will be accrued by new customers, even though we have found a significant relationship between these two variables. If the model is to be used to predict future amounts of credit charges, then the usefulness of including the number of hours per week the consumer watches television is best evaluated by measuring the accuracy of predictions for observations not included in the sample data used to construct the model. We demonstrate this procedure in the next subsection.

The use of out-of-sample data is common in data mining applications and is covered in detail in Chapter 10.

Model Selection

As we discussed in Section 8.8, various methods for identifying which independent variables to include in a linear regression model consider the *p* values of these variables in iterative procedures that sequentially add and/or remove variables. However, when dealing with a sufficiently large sample, the *p* value of virtually every independent variable will be small, and so variable selection procedures may suggest models with most or all the variables. Therefore, when dealing with large samples, it is often more difficult to discern the most appropriate model.

If developing a linear regression model for explanatory purposes, the practical significance of the estimated coefficients should be considered when interpreting the model and considering which variables to keep in the model. If developing a linear regression model

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to make future predictions, the selection of the independent variables to include in the linear regression model should be based on the predictive accuracy of the model on observations that have not been used to train the model.

Let us revisit the example of a credit card company with a data set of customer records containing information on the customer's annual household income, number of years of post–high school education, number of members of the customer's household, and the credit card charges accrued. Recall that the file *largecredit* contains these data, split into a training set of 3,000 observations and a validation set of 2,000 observations.

To predict annual credit card charges for its new applicants, the company is considering two models:

- Model A—The dependent variable is credit card charges accrued by a customer in the data set over the past year (y), and the independent variables are the customer's annual household income (x_1), number of household members (x_2), and number of years of post–high school education (x_3). Figure 8.37 summarizes Model A estimated using the 3,000 observations of the training set.
- Model B—The dependent variable is credit card charges accrued by a customer in the data set over the past year (y), and the independent variables are the customer's annual household income (x_1), number of household members (x_2), number of years of post–high school education (x_3), and number of hours per week spent watching television (x_4). Figure 8.38 summarizes Model B estimated using the 3,000 observations of the training set.

Now, we would like to compare these models based on their predictive accuracy on the 2,000 observations in the validation set. For the first observation in the validation set (account number 18572870), Model A predicts annual charges of

$$\hat{y}_1^A = 2119.60 + 121.33(50.2) + 528.10(5) - 525.36(1) = \$10{,}315.93$$

Alternatively, Model B predicts annual charges of

$$\hat{y}_1^B = 1712.55 + 121.61(50.2) + 531.21(5) - 539.89(1) + 12.55(4) = \$9{,}983.92$$

Account number 18572870 has actual annual charges of \$5,472.51, so Model A's prediction of the first observation has a squared error of $(5{,}472.51 - 10{,}315.93)^2 = 23{,}458{,}721$ and Model B's prediction of the first observation has a squared error of $(5{,}472.51 - 9{,}983.92)^2 = 20{,}352{,}797$. Repeating these predictions and error calculations for each of the 2,000 observations in the validation set, Figure 8.39 shows that the sum of squared errors for Model A is 47,392,009,111 and the sum of squared errors for Model B is 47,409,404,281. Therefore, Model A's predictions are slightly more accurate than Model B's predictions on the validation set, as measured by squared error. Although the p value of Hours Per Week Watching Television in Model B is relatively small, these results suggest that adding it to Model A does not improve the accuracy of predictions.

8.10 Prediction with Linear Regression

To illustrate how a linear regression model can be used to make predictions about new observations and support decision making, let us again consider the Butler Trucking Company. Recall from Section 8.4 that the multiple linear regression equation based on the 300 past routes using Miles (x_1) and Deliveries (x_2) as the independent variables to estimate travel time (y) for a driving assignment is

$$\hat{y} = 0.1273 + 0.0672x_1 + 0.6900x_2 \quad \textbf{(8.22)}$$

As described by the first three columns of Table 8.5, Butler has 10 new observations corresponding to upcoming routes for which they have estimated the miles to be driven and number of deliveries. The point estimates for the travel time for each of these 10 upcoming

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 8.39 Predictive Accuracy on *LargeCredit* Validation Set

	A	B	C	D	E	F	G	H	I	J	K	L
1								Model A (3 Variable)			Model B (4 Variable)	
2	Observation	Account Number	Annual Income ($1000)	Household Size	Years of Post-High School Education	Hours Per Week Watching Television	Annual Charges ($)	Prediction	Squared Error		Prediction	Squared Error
3	3001	18572870	50.2	5.0	1.0	4.0	5472.51	10,315.93	23,458,721.44		9,983.92	20,352,797.45
4	3002	10135558	39.6	2.0	4.0	15.0	3968.42	5,839.37	3,500,437.29		5,619.76	2,726,908.40
5	3003	23467852	88.8	4.0	1.0	19.0	11382.63	14,471.50	9,541,090.64		14,335.21	8,717,710.16
6	3004	2221007	101.2	6.0	2.0	54.0	16827.83	16,496.93	109,493.08		16,805.10	516.60
7	3005	23024579	52.0	5.0	2.0	19.0	13175.27	9,998.98	10,088,816.14		9,851.26	11,049,033.20
8	3006	5534868	100.8	5.0	4.0	50.0	20292.73	14,849.58	29,627,894.64		15,095.37	27,012,585.44
9	3007	19704869	70.6	3.0	0.0	49.0	6230.89	12,270.40	36,475,622.43		12,507.04	39,390,082.33
10	3008	9388137	88.9	8.0	2.0	41.0	18914.62	16,060.67	8,145,037.30		16,208.53	7,322,943.68
11	3009	23883625	89.4	5.0	2.0	29.0	14362.00	14,537.04	30,638.65		14,525.07	26,592.07
1991	4989	6776616	87.6	3.0	2.0	59.0	20541.21	13,262.43	52,980,632.57		13,620.30	47,899,053.37
1992	4990	8695442	47.3	8.0	1.0	10.0	17011.33	11,548.35	29,844,173.12		11,300.19	32,617,082.88
1993	4991	5888985	82.4	4.0	1.0	48.0	9416.69	13,694.93	18,303,332.35		13,920.89	20,287,829.66
1994	4992	12243467	43.2	5.0	1.0	16.0	3101.00	9,466.56	40,520,368.82		9,283.25	38,220,269.20
1995	4993	28297658	49.9	5.0	0.0	48.0	14538.99	10,814.89	13,868,933.92		11,039.55	12,246,101.91
1996	4994	4605783	36.7	3.0	2.0	19.0	12620.39	7,086.30	30,626,125.63		6,928.17	32,401,368.92
1997	4995	21430617	54.9	1.0	5.0	27.0	3755.45	6,632.39	8,276,755.45		6,559.99	7,865,464.34
1998	4996	3080483	84.4	4.0	5.0	23.0	13018.42	11,796.17	1,493,897.69		11,690.98	1,762,090.04
1999	4997	8089356	41.6	2.0	4.0	14.0	8740.70	6,082.04	7,068,459.72		5,850.43	8,353,673.98
2000	4998	14223252	51.0	7.0	5.0	4.0	0.00	9,327.76	87,007,165.68		8,984.30	80,717,567.08
2001	4999	8048637	39.0	7.0	1.0	5.0	360.73	10,013.14	93,168,998.76		9,696.84	87,162,964.44
2002	5000	27638369	39.0	5.0	2.0	19.0	1554.11	8,421.58	47,162,147.49		8,270.30	45,107,267.97
2003												
2004							SSE:		47,392,009,111			47,409,404,281

Table 8.5 Predicted Values and 95% Confidence Intervals and Prediction Intervals for 10 New Butler Trucking Routes

Assignment	Miles	Deliveries	Predicted Value	95% CI Half-Width(±)	95% PI Half-Width(±)
301	105	3	9.25	0.193	1.645
302	60	4	6.92	0.112	1.637
303	95	5	9.96	0.173	1.642
304	100	1	7.54	0.225	1.649
305	40	3	4.88	0.177	1.643
306	80	3	7.57	0.108	1.637
307	65	4	7.25	0.103	1.637
308	55	3	5.89	0.124	1.638
309	95	2	7.89	0.175	1.643
310	95	3	8.58	0.154	1.641

routes can be obtained by substituting the miles driven and number of deliveries into equation (8.22). For example, the predicted travel time for Assignment 301 is

$$\hat{y}_{301} = 0.1273 + 0.0672(105) + 0.6900(3) = 9.25$$

In addition to the point estimate, there are two types of interval estimates associated with the linear regression equation. A confidence interval is an interval estimate of the mean y value given values of the independent variables. A **prediction interval** is an interval estimate of an individual y value given values of the independent variables.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The general form for the confidence interval on the mean y value given values of $x_1, x_2, \ldots, x_q$ is

$$\hat{y} \pm t_{\alpha/2} s_{\hat{y}} \tag{8.23}$$

where $\hat{y}$ is the point estimate of the mean y value provided by the linear regression equation, $s_{\hat{y}}$ is the estimated standard deviation of $\hat{y}$, and $t_{a/2}$ is a multiplier term based on the sample size and specified $100(1-\alpha)\%$ confidence level of the interval. More specifically, $t_{a/2}$ is the t value that provides an area of $\alpha/2$ in the upper tail of a t distribution with $n - q - 1$ degrees of freedom. In general, the calculation of the confidence interval in equation (8.23) uses matrix algebra and requires the use of specialized statistical software.

The prediction interval on the individual y value given values of $x_1, x_2, \ldots, x_q$ is

$$\hat{y} \pm t_{\alpha/2} \sqrt{s_{\hat{y}}^2 + \frac{SSE}{n - q - 1}} \tag{8.24}$$

where $\hat{y}$ is the point estimate of the individual y value provided by the linear regression equation, $s_{\hat{y}}^2$ is the estimated variance of $\hat{y}$, and $t_{a/2}$ is the t value that provides an area of $\alpha/2$ in the upper tail of a t distribution with $n - q - 1$ degrees of freedom. In the term $SSE/(n - q - 1)$, n is the number of observations in the sample, q is the number of independent variables in the linear regression model, and SSE is the sum of squares due to error as defined by equation (8.5). In general, the calculation of the prediction interval in equation (8.24) uses matrix algebra and requires the use of specialized statistical software.

Statistical software such as R can be used to compute the confidence interval and prediction intervals on linear regression output.

In the Butler Trucking problem, the 95% confidence interval is an interval estimate of the mean travel time for a route assignment with the given values of Miles and Deliveries. This is the appropriate interval estimate if we are interested in estimating the mean travel time for all route assignments with specified mileage and number of deliveries. This confidence interval estimates the variability in the mean travel time.

In the Butler Trucking problem, the 95% prediction interval is an interval estimate on the prediction of travel time for an individual route assignment with the given values of Miles and Deliveries. This is the appropriate interval estimate if we are interested in predicting the travel time for an individual route assignment with the specified mileage and number of deliveries. This prediction interval estimates the variability inherent in a single route's travel time.

To illustrate, consider the first observation (Assignment 301) in Table 8.5 with 105 miles and 3 deliveries. For all 105-mile routes with 3 deliveries, a 95% confidence interval on the mean travel time is 9.25 ± 0.193. That is, we are 95% confident that the true population mean travel time for 105-mile routes with 3 deliveries is between 9.06 and 9.44 hours.

Now suppose Butler Trucking is interested in predicting the travel time for a specific upcoming route assignment covering 105 miles and 3 deliveries. The best prediction for this route's travel time is still 9.25 hours, as provided by the linear regression equation. However, a 95% prediction interval for this travel time prediction is 9.25 ± 1.645. That is, we are 95% confident that the travel time for a single 105-mile route with 3 deliveries will be between 7.61 and 10.90 hours.

Note that the 95% prediction interval for the travel time of a single route assignment with 105 miles and 3 deliveries is wider than the 95% confidence interval for the mean travel time of all route assignments with 105 miles and 3 deliveries. The difference reflects the fact that we are able to estimate the mean y value of a group of observations with the same specified values of the independent variables more precisely than we can predict an individual y value of a single observation with specified values of the independent variables. Comparing equation (8.23) to equation (8.24), we observe the reason for the difference in the width of the confidence interval and the prediction interval. Just as the confidence interval, the prediction interval calculation includes an $s_{\hat{y}}$ term to account for the variability in estimating the mean value of y, but it also includes an additional term $SSE/(n - q - 1)$ which accounts for the variability in individual values of y about its mean value.

Finally, we point out that the width of the prediction (and confidence) intervals for the linear regression model's point estimate are not the same for each observation. Instead, the width of the interval depends on the corresponding values of the independent variables. Confidence intervals and prediction intervals are the narrowest when the values of the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

independent variables, $x_1, x_2, \ldots, x_q$, are closest to their respective means, $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_q$. For the 300 observations on which the linear regression equation model is based, the mean miles driven for a route assignment is 70.7 miles and the mean number of deliveries for a route assignment is 3.5. Assignment 307 has the mileage (65) and number of deliveries (4) that are closest to these means and correspondingly has the narrowest confidence and prediction interval. Conversely, Assignment 304 has the widest confidence and prediction intervals because it has the mileage (100) and number of deliveries (1) that are the farthest from the mean mileage and mean number of deliveries in the data.

Summary

In this chapter we showed how linear regression analysis can be used to determine how a dependent variable y is related to an independent variable x. In simple linear regression, the model is $y = \beta_0 + \beta_1 x_1 + \varepsilon$. We use sample data and the least squares method to develop the estimated linear regression equation $\hat{y} = b_0 + b_1 x_1$. In effect, b_0 and b_1 are the sample statistics used to estimate the unknown model parameters.

The coefficient of determination was presented as a measure of the goodness of fit for the estimated linear regression equation; it can be interpreted as the proportion of the variation in the sample values of the dependent variable y that can be explained by the estimated linear regression equation. We then extended our discussion to include multiple independent variables and reviewed how to use Excel to find the estimated multiple linear regression equation $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_q x_q$, and we considered the interpretations of the parameter estimates in multiple linear regression and the ramifications of multicollinearity.

The assumptions related to the linear regression model and its associated error term ϵ were discussed. We reviewed the t test for determining whether there is a statistically significant relationship between the dependent variable and an individual independent variable given the other independent variables in the linear regression model. We also showed how to use Excel to develop confidence interval estimates of the linear parameters $\beta_0, \beta_1, \beta_2 \ldots, \beta_q$.

We showed how to incorporate categorical independent variables into a linear regression model through the use of dummy variables, and we discussed a variety of ways to use multiple regression to fit nonlinear relationships between independent variables and the dependent variable. We discussed various automated procedures for selecting independent variables to include in a linear regression model and the problem of overfitting a linear regression model.

We discussed the implications of big data on linear regression analysis. Specifically, we considered the impact of very large samples on linear regression inference and demonstrated the use of holdout data to evaluate candidate linear regression models. We concluded by presenting the concepts of confidence intervals and prediction intervals related to point estimates produced by the linear regression model.

Glossary

Backward elimination An iterative variable selection procedure that starts with a model with all independent variables and considers removing an independent variable at each step.
Best subsets A variable selection procedure that constructs and compares all possible models with up to a specified number of independent variables.
Coefficient of determination A measure of the goodness of fit of the estimated linear regression equation. It can be interpreted as the proportion of the variability in the dependent variable y that is explained by the estimated linear regression equation.
Confidence interval An estimate of a population parameter that provides an interval believed to contain the value of the parameter at some level of confidence.
Confidence level An indication of how frequently interval estimates based on samples of the same size taken from the same population using identical sampling techniques will contain the true value of the parameter we are estimating.
Cross-validation Assessment of the performance of a model on data other than the data that were used to generate the model.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Dependent variable The variable that is being predicted or explained. It is denoted by y and is often referred to as the response variable or target variable.

Dummy variable A variable used to model the effect of categorical independent variables in a linear regression model; generally takes only the value zero or one.

Estimated linear regression The estimate of the linear regression equation developed from sample data by using the least squares method. The estimated multiple linear regression equation is $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_qx_q$.

Experimental region The range of values for the independent variables $x_1, x_2, \ldots, x_q$ for the data that are used to estimate the linear regression model.

Extrapolation Prediction of the mean value of the dependent variable y for values of the independent variables $x_1, x_2, \ldots, x_q$ that are outside the experimental range.

Forward selection An iterative variable selection procedure that starts with a model with no variables and considers adding an independent variable at each step.

Holdout method Method of cross-validation in which sample data are randomly divided into mutually exclusive and collectively exhaustive sets, then one set is used to build the candidate models and the other set is used to compare model performances and ultimately select a model.

Hypothesis testing The process of making a conjecture about the value of a population parameter, collecting sample data that can be used to assess this conjecture, measuring the strength of the evidence against the conjecture that is provided by the sample, and using these results to draw a conclusion about the conjecture.

Independent variable(s) The variable(s) used for predicting or explaining values of the dependent variable. It is denoted by x and is often referred to as a predictor variable or a feature.

Interaction Regression modeling technique used when the relationship between the dependent variable and one independent variable is different at different values of a second independent variable.

Interval estimation The use of sample data to calculate a range of values that is believed to include the unknown value of a population parameter.

Knot The prespecified value of the independent variable at which its relationship with the dependent variable changes in a piecewise linear regression model; also called the breakpoint or the joint.

Least squares method A procedure for using sample data to find the estimated linear regression equation.

Linear regression model The equation that describes how the dependent variable y is related to an independent variable x and an error term; the *multiple linear regression model* is $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_qx_q + \varepsilon$.

Multicollinearity The degree of correlation among independent variables in a linear regression model.

Multiple linear regression Linear regression analysis involving one dependent variable and more than one independent variable.

Overfitting Fitting a model too closely to sample data, resulting in a model that does not accurately reflect the population.

***p* value** The probability that a random sample of the same size collected from the same population using the same procedure will yield stronger evidence against a hypothesis than the evidence in the sample data given that the hypothesis is actually true.

Piecewise linear regression model Regression model in which one linear relationship between the independent and dependent variables is fit for values of the independent variable below a prespecified value of the independent variable, a different linear relationship between the independent and dependent variables is fit for values of the independent variable above the prespecified value of the independent variable, and the two linear regression models have the same estimated value of the dependent variable (i.e., are joined) at the prespecified value of the independent variable.

Point estimator A single value used as an estimate of the corresponding population parameter.

Population parameter A measurable factor that defines a characteristic of a population, process, or system.

Prediction interval An interval estimate of the prediction of an individual y value given values of the independent variables.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Quadratic regression model Regression model in which a nonlinear relationship between the independent and dependent variables is fit by including the independent variable and the square of the independent variable in the model: $\hat{y} = b_o + b_1 x_1 + b_2 x_1^2$; also referred to as a second-order polynomial model.
Regression A prediction task in which the variable to be predicted is quantitative.
Residual The difference between the observed value of the dependent variable and the value predicted using the estimated regression equation; for the ith observation, the ith residual is $y_i - \hat{y}_i$.
Simple linear regression A linear regression model involving one dependent variable and one independent variable.
Statistical inference The process of making estimates and drawing conclusions about one or more characteristics of a population (the value of one or more parameters) through analysis of sample data drawn from the population.
Stepwise selection An iterative variable selection procedure that considers adding an independent variable and removing an independent variable at each step.
***t* test** Statistical test based on the Student's t probability distribution that can be used to test the hypothesis that a linear regression parameter β_j is zero; if this hypothesis is rejected, we conclude that there is a relationship between the jth independent variable and the dependent variable.
Training set The data set used to build the candidate models.
Validation set The data set used to compare model forecasts and ultimately pick a model for predicting values of the dependent variable.

Problems

1. **Price and Weight of Bicycles.** *Bicycling World*, a magazine devoted to cycling, reviews hundreds of bicycles throughout the year. Its Road-Race category contains reviews of bicycles used by riders primarily interested in racing. One of the most important factors in selecting a bicycle for racing is its weight. The following data show the weight (pounds) and price ($) for 10 racing bicycles reviewed by the magazine. **LO 1, 3, 4, 5**

DATA*file* bicyclingworld

Model	Weight (lb)	Price ($)
Fierro 7B	17.9	2,200
HX 5000	16.2	6,350
Durbin Ultralight	15.0	8,470
Schmidt	16.0	6,300
WSilton Advanced	17.3	4,100
bicyclette vélo	13.2	8,700
Supremo Team	16.3	6,100
XTC Racer	17.2	2,680
D'Onofrio Pro	17.7	3,500
Americana #6	14.2	8,100

   a. Develop a scatter chart with weight as the independent variable. What does the scatter chart indicate about the relationship between the weight and price of these bicycles?
   b. Use the data to develop an estimated linear regression equation that could be used to estimate the price for a bicycle, given its weight. What is the estimated linear regression model?
   c. Test whether each of the parameters β_0 and β_1 is equal to zero at a 0.05 level of significance. What are the correct interpretations of the estimated parameters? Are these interpretations reasonable?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

d. How much of the variation in the prices of the bicycles in the sample does the linear regression model you estimated in part (b) explain?
e. The manufacturers of the D'Onofrio Pro plan to introduce the 15-lb D'Onofrio Elite bicycle later this year. Use the linear regression model you estimated in part (a) to predict the price of the D'Ononfrio Elite.
f. The owner of Michele's Bikes of Nesika Beach, Oregon is trying to decide in advance whether to make room for the D'Onofrio Elite bicycle in its inventory. The owner is convinced that they will not be able to sell the D'Onofrio Elite for more than $7,000, and so they will not make room in their inventory for the bicycle unless its estimated price is less than $7,000. Under this condition and using the linear regression model you estimated in part (a), what decision should the owner of Michele's Bikes make?

2. **Production Rate and Quality Control.** In a manufacturing process the assembly line speed (feet per minute) was thought to affect the number of defective parts found during the inspection process. To test this theory, managers devised a situation in which the same batch of parts was inspected visually at a variety of line speeds. They collected the following data. **LO 1, 3, 4, 5**

DATA*file*
linespeed

Line Speed (ft/min)	No. of Defective Parts Found
20	21
20	19
40	15
30	16
60	14
40	17

a. Develop a scatter chart with line speed as the independent variable. What does the scatter chart indicate about the relationship between line speed and the number of defective parts found?
b. Use the data to develop an estimated linear regression equation that could be used to predict the number of defective parts found, given the line speed. What is the estimated linear regression model?
c. Test whether each of the parameters β_0 and β_1 is equal to zero at a 0.01 level of significance. What are the correct interpretations of the estimated parameters? Are these interpretations reasonable?
d. How much of the variation in the number of defective parts found for the sample data does the model you estimated in part (b) explain?

DATA*file*
jensens

3. **Machine Maintenance.** Jensen's Tire & Auto is deciding whether to purchase a maintenance contract for its recently installed computer wheel alignment and balancing machines. Managers feel that maintenance expense should be related to usage, and they collect information on weekly usage (hours) and annual maintenance expense (in hundreds of dollars) from 300 Jensen's Tire & Auto locations throughout the nation. These data are available in the file *jensens*. **LO 1, 3, 4, 5, 9**
a. Develop a scatter chart with weekly usage hours as the independent variable. What does the scatter chart indicate about the relationship between weekly usage and annual maintenance expense?
b. Use the data to develop an estimated linear regression equation that could be used to predict the annual maintenance expense for a given number of hours of weekly usage. What is the estimated linear regression model?
c. Test whether each of the parameters β_0 and β_1 is equal to zero at a 0.05 level of significance. What are the correct interpretations of the estimated parameters? Are these interpretations reasonable?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

d. How much of the variation in the sample values of annual maintenance expense does the model you estimated in part (b) explain?
e. If the maintenance contract costs $3,000 per year, would you recommend purchasing it? Why or why not?

DATA*file* absentee

4. **Absenteeism and Location.** A sociologist was hired by a large city hospital to investigate the relationship between the number of unauthorized days that employees are absent per year and the distance (miles) between home and work for the employees. Data on these variables were collected from a sample of 100 employees, and the data are available in the file *absentee*. **LO 1, 3, 4, 5, 6**
   a. Develop a scatter chart with distance to work as the independent variable. Does a linear relationship appear reasonable? Explain.
   b. Use the data to develop an estimated linear regression equation that could be used to predict the number of days absent given the distance to work. What is the estimated linear regression model?
   c. What is the 99% confidence interval for the parameter β_1? Based on this interval, what conclusion can you make about the hypotheses that the parameter β_1 is equal to zero?
   d. What is the 99% confidence interval for the parameter β_0? Based on this interval, what conclusion can you make about the hypotheses that the parameter β_0 is equal to zero?
   e. How much of the variation in the sample values of number of days absent does the model you estimated in part (b) explain?

DATA*file* busagecost

5. **Bus Maintenance.** The regional transit authority for a major metropolitan area wants to determine whether there is a relationship between the age of a bus and the annual maintenance cost. A sample of 73 buses resulted in the data provided in the file *busagecost*. **LO 1, 3, 4, 5**
   a. Develop a scatter chart with age of bus as the independent variable. What does the scatter chart indicate about the relationship between age of a bus and the annual maintenance cost?
   b. Use the data to develop an estimated linear regression equation that could be used to predict the annual maintenance cost given the age of the bus. What is the estimated linear regression model?
   c. Test whether each of the parameters β_0 and β_1 is equal to zero at a 0.05 level of significance. What are the correct interpretations of the estimated parameters? Are these interpretations reasonable?
   d. How much of the variation in the sample values of annual maintenance cost does the model you estimated in part (b) explain?
   e. What do you predict the annual maintenance cost to be for a 3.5-year-old bus?

DATA*file* marketinghrs

6. **Studying and Grades.** A marketing professor at Givens College is interested in the relationship between hours spent studying and total points earned in a course. Data collected on 156 students who took the course last semester are provided in the file *marketinghrs*. **LO 1, 3, 4, 5, 9**.
   a. Develop a scatter chart for these data. What does the scatter chart indicate about the relationship between total points earned and hours spent studying?
   b. Develop an estimated linear regression equation showing how total points earned is related to hours spent studying. What is the estimated linear regression model?
   c. Test whether each of the parameters β_0 and β_1 is equal to zero at a 0.01 level of significance. What are the correct interpretations of the estimated parameters? Are these interpretations reasonable?
   d. How much of the variation in the sample values of total point earned does the model you estimated in part (b) explain?
   e. Mark Sweeney spent 95 hours studying. Use the linear regression model you estimated in part (b) to predict the total points Mark earned.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

f. Mark Sweeney wants to receive a letter grade of A for this course, and Mark needs to earn at least 90 points to do so. Based on the linear regression equation developed in part (b), how many estimated hours should Mark study to receive a letter grade of A for this course?

DATA*file* closingprices

7. **Stock Market Performance.** The Dow Jones Industrial Average (DJIA) and the Standard & Poor's 500 (S&P 500) indexes are used as measures of overall movement in the stock market. The DJIA is based on the price movements of 30 large companies; the S&P 500 is an index composed of 500 stocks. Some say the S&P 500 is a better measure of stock market performance because it is broader based. The closing price for the DJIA and the S&P 500 for 144 days in 2022 are provided in the file *closingprices*. (*The Wall Street Journal* web site). **LO 1, 3, 4, 5, 6, 9**
   a. Develop a scatter chart for these data with DJIA as the independent variable. What does the scatter chart indicate about the relationship between DJIA and S&P 500?
   b. Develop an estimated linear regression equation showing how S&P 500 is related to DJIA. What is the estimated linear regression model?
   c. What is the 95% confidence interval for the parameter β_1? Based on this interval, what conclusion can you make about the hypotheses that the parameter β_1 is equal to zero?
   d. What is the 95% confidence interval for the parameter β_0? Based on this interval, what conclusion can you make about the hypotheses that the parameter β_0 is equal to zero?
   e. How much of the variation in the sample values of S&P 500 does the model estimated in part (b) explain?
   f. Suppose that the closing price for the DJIA is 38,500. Estimate the closing price for the S&P 500.
   g. Should we be concerned that the DJIA value of 38,500 used to predict the S&P 500 value in part (f) is beyond the range of the DJIA used to develop the estimated linear regression equation?

DATA*file* camrys

8. **Used Car Mileage and Price.** The Toyota Camry is one of the best-selling cars in North America. The cost of a previously owned Camry depends on many factors, including the model year, mileage, and condition. To investigate the relationship between the car's mileage and the sales price for Camrys, a data scientist collects the the mileage and sale price for 124 recent sales of Toyota Camrys. The data are available in the file *camrys*. **LO 1, 3, 4, 5, 9**
   a. Develop a scatter chart for these data with miles as the independent variable. What does the scatter chart indicate about the relationship between price and miles?
   b. Develop an estimated linear regression equation showing how price is related to miles. What is the estimated linear regression model?
   c. Test whether each of the parameters β_0 and β_1 is equal to zero at a 0.01 level of significance. What are the correct interpretations of the estimated parameters? Are these interpretations reasonable?
   d. How much of the variation in the sample values of price does the model estimated in part (b) explain?
   e. For the model estimated in part (b), calculate the predicted price and residual for each automobile in the data. Identify the two automobiles that were the biggest bargains.
   f. Suppose that you are considering purchasing a previously owned Camry that has been driven 110,000 miles. Use the estimated linear regression equation developed in part (b) to predict the price for this car. Is this the price you would offer the seller?

9. **Weekly Theater Revenue.** Southerly Showtime Movie Theaters, Inc. owns and operates a chain of cinemas in several markets in the southern United States. The owners would like to estimate weekly gross revenue as a function of advertising expenditures. Data for a sample of eight markets for a recent week follow. **LO 1, 2, 3, 4, 5**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA file
southerlyshowtime

Market	Weekly Gross Revenue ($100s)	Television Advertising ($100s)	Newspaper Advertising ($100s)
Mobile	101.3	5.0	1.5
Shreveport	51.9	3.0	3.0
Jackson	74.8	4.0	1.5
Birmingham	126.2	4.3	4.3
Little Rock	137.8	3.6	4.0
Biloxi	101.4	3.5	2.3
New Orleans	237.8	5.0	8.4
Baton Rouge	219.6	6.9	5.8

a. Develop an estimated linear regression equation with the amount of television advertising as the independent variable. Test for a significant relationship between television advertising and weekly gross revenue at the 0.05 level of significance. What is the interpretation of this relationship?
b. How much of the variation in the sample values of weekly gross revenue does the model in part (a) explain?
c. Develop an estimated linear regression equation with both television advertising and newspaper advertising as the independent variables. Test whether each of the parameters β_0, β_1, and β_2 is equal to zero at a 0.05 level of significance. What are the correct interpretations of the estimated parameters? Are these interpretations reasonable?
d. How much of the variation in the sample values of weekly gross revenue does the model in part (c) explain?
e. Given the results in parts (a) and (c), what should your next step be? Explain.
f. What are the managerial implications of these results?

10. **Ratings of Beachfront Boutique Hotels.** *Resorts & Spas*, a magazine devoted to upscale vacations and accommodations, published its Reader's Choice List of the top 20 independent beachfront boutique hotels in the world. The data shown are the scores received by these hotels based on the results from *Resorts & Spas'* annual Readers' Choice Survey. Each score represents the percentage of respondents who rated a hotel as excellent or very good on one of three criteria (comfort, amenities, and in-house dining). An overall score was also reported and used to rank the hotels. The highest ranked hotel, the Muri Beach Odyssey, has an overall score of 94.3, the highest component of which is 97.7 for in-house dining. **LO 2, 4, 5, 9**

DATA file
beachfronthotels

Hotel	Overall	Comfort	Amenities	In-House Dining
Muri Beach Odyssey	94.3	94.5	90.8	97.7
Pattaya Resort	92.9	96.6	84.1	96.6
Sojourner's Respite	92.8	99.9	100.0	88.4
Spa Carribe	91.2	88.5	94.7	97.0
Penang Resort and Spa	90.4	95.0	87.8	91.1
Mokihana Ho-kele	90.2	92.4	82.0	98.7
Theo's of Cape Town	90.1	95.9	86.2	91.9
Cap d'Agde Resort	89.8	92.5	92.5	88.8
Spirit of Mykonos	89.3	94.6	85.8	90.7
Turismo del Mar	89.1	90.5	83.2	90.4
Hotel Iguana	89.1	90.8	81.9	88.5
Sidi Abdel Rahman Palace	89.0	93.0	93.0	89.6

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Hotel	Overall	Comfort	Amenities	In-House Dining
Sainte-Maxime Quarters	88.6	92.5	78.2	91.2
Rotorua Inn	87.1	93.0	91.6	73.5
Club Lapu-Lapu	87.1	90.9	74.9	89.6
Terracina Retreat	86.5	94.3	78.0	91.5
Hacienda Punta Barco	86.1	95.4	77.3	90.8
Rendezvous Kolocep	86.0	94.8	76.4	91.4
Cabo de Gata Vista	86.0	92.0	72.2	89.2
Sanya Deluxe	85.1	93.4	77.3	91.8

a. Determine the estimated multiple linear regression equation that can be used to predict the overall score given the scores for comfort, amenities, and in-house dining.
b. Use the *t* test to determine the significance of each independent variable. What is the conclusion for each test at the 0.01 level of significance?
c. Remove all independent variables that are not significant at the 0.01 level of significance from the estimated linear regression equation. What is your recommended estimated linear regression equation?
d. Suppose *Resorts & Spas* has decided to recommend each of the independent beachfront boutiques in its data that achieves an estimated overall score over 90. Use the linear regression equation developed in part (c) to determine which of the independent beachfront boutiques will receive a recommendation from *Resorts & Spas*.

11. **Trading Stocks Electronically.** The American Association of Individual Investors (AAII) On-Line Discount Broker Survey polls members on their experiences with electronic trades handled by discount brokers. As part of the survey, members were asked to rate their satisfaction with the trade price and the speed of execution, as well as provide an overall satisfaction rating. Possible responses (scores) were no opinion (0), unsatisfied (1), somewhat satisfied (2), satisfied (3), and very satisfied (4). For each broker, summary scores were computed by computing a weighted average of the scores provided by each respondent. A portion of the survey results follow (AAII web site). **LO 2, 4, 5, 9**

DATA*file*
broker

Brokerage	Satisfaction with Trade Price	Satisfaction with Speed of Execution	Overall Satisfaction with Electronic Trades
Scottrade, Inc.	3.4	3.4	3.5
Charles Schwab	3.2	3.3	3.4
Fidelity Brokerage Services	3.1	3.4	3.9
TD Ameritrade	2.9	3.6	3.7
E*Trade Financial	2.9	3.2	2.9
(Not listed)	2.5	3.2	2.7
Vanguard Brokerage Services	2.6	3.8	2.8
USAA Brokerage Services	2.4	3.8	3.6
Thinkorswim	2.6	2.6	2.6
Wells Fargo Investments	2.3	2.7	2.3
Interactive Brokers	3.7	4.0	4.0
Zecco.com	2.5	2.5	2.5
Firstrade Securities	3.0	3.0	4.0
Banc of America Investment Services	4.0	1.0	2.0

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Develop an estimated linear regression equation using trade price and speed of execution to predict overall satisfaction with the broker. Interpret the coefficient of determination.
b. Use the *t* test to determine the significance of each independent variable. What are your conclusions at the 0.05 level of significance?
c. Interpret the estimated parameters. Are the relationships indicated by these estimates what you would expect?
d. Finger Lakes Investments has developed a new electronic trading system and would like to predict overall customer satisfaction assuming they can provide satisfactory service levels (3) for both trade price and speed of execution. Use the estimated linear regression equation developed in part (a) to predict overall satisfaction level for Finger Lakes Investments if they can achieve these performance levels.
e. What concerns (if any) do you have with regard to the possible responses the respondents could select on the survey.

12. **NFL Winning Percentage.** The National Football League (NFL) records a variety of performance data for individuals and teams. To investigate the importance of passing on the percentage of games won by a team, the following data show the conference (Conf), average number of passing yards per attempt (Yds/Att), the number of interceptions thrown per attempt (Int/Att), and the percentage of games won (Win%) for a random sample of 16 NFL teams for the 2011 season (NFL web site). **LO 1, 2, 3, 9**

DATA*file*
nflpassing

Team	Conf	Yds/Att	Int/Att	Win%
Arizona Cardinals	NFC	6.5	0.042	50.0
Atlanta Falcons	NFC	7.1	0.022	62.5
Carolina Panthers	NFC	7.4	0.033	37.5
Cincinnati Bengals	AFC	6.2	0.026	56.3
Detroit Lions	NFC	7.2	0.024	62.5
Green Bay Packers	NFC	8.9	0.014	93.8
Houston Texans	AFC	7.5	0.019	62.5
Indianapolis Colts	AFC	5.6	0.026	12.5
Jacksonville Jaguars	AFC	4.6	0.032	31.3
Minnesota Vikings	NFC	5.8	0.033	18.
New England Patriots	AFC	8.3	0.020	81.3
New Orleans Saints	NFC	8.1	0.021	81.3
Oakland Raiders	AFC	7.6	0.044	50.0
San Francisco 49ers	NFC	6.5	0.011	81.3
Tennessee Titans	AFC	6.7	0.024	56.3
Washington Commanders	NFC	6.4	0.041	31.3

a. Develop the estimated linear regression equation that could be used to predict the percentage of games won, given the average number of passing yards per attempt. What proportion of variation in the sample values of proportion of games won does this model explain?
b. Develop the estimated linear regression equation that could be used to predict the percentage of games won, given the number of interceptions thrown per attempt. What proportion of variation in the sample values of proportion of games won does this model explain?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Develop the estimated linear regression equation that could be used to predict the percentage of games won, given the average number of passing yards per attempt and the number of interceptions thrown per attempt. What proportion of variation in the sample values of proportion of games won does this model explain?

d. The average number of passing yards per attempt for the Kansas City Chiefs during the 2011 season was 6.2, and the team's number of interceptions thrown per attempt was 0.036. Use the estimated linear regression equation developed in part (c) to predict the percentage of games won by the Kansas City Chiefs during the 2011 season. Compare your prediction to the actual percentage of games won by the Kansas City Chiefs. (*Note*: For the 2011 season, the Kansas City Chiefs' record was 7 wins and 9 losses.)

e. Did the estimated linear regression equation that uses only the average number of passing yards per attempt as the independent variable to predict the percentage of games won provide a good fit?

DATA*file*
repairs

13. **Water Filtration System Maintenance.** Johnson Filtration, Inc. provides maintenance service for water filtration systems throughout Southern Florida. Customers contact Johnson with requests for maintenance service on their water filtration systems. To estimate the service time and the service cost, Johnson's managers want to predict the repair time necessary for each maintenance request. Hence, repair time in hours is the dependent variable. Repair time is believed to be related to three factors: the number of months since the last maintenance service, the type of repair problem (mechanical or electrical), and the repairperson who performs the repair (Donna Newton or Bob Jones). Data for a sample of 50 service calls are reported in the file *repairs*.
LO 1, 2, 3, 4, 5, 7:

a. Develop the simple linear regression equation to predict repair time given the number of months since the last maintenance service, and use the results to test the hypothesis that no relationship exists between repair time and the number of months since the last maintenance service at the 0.05 level of significance. What is the interpretation of this relationship? What does the coefficient of determination tell you about this model?

b. Using the simple linear regression model developed in part (a), calculate the predicted repair time and residual for each of the 50 repairs in the data and append the original data of 50 observations with these two new columns. Sort the data in ascending order by value of the residual. Do you see any pattern in the residuals with respect to the type of repair or repairperson? Do these results suggest any potential modifications to your simple linear regression model?

c. Create a new dummy variable called Newton that is equal to zero if the repair person is Bob Jones and one if the repair person is Donna Newton. Develop the multiple linear regression equation to predict repair time, given the number of months since the last maintenance service and the repairperson. What are the interpretations of the estimated parameters? What does the coefficient of determination tell you about this model?

d. Create a new dummy variable called Electrical that is equal to zero if the type of repair is mechanical and one if the type of repair is electrical. Develop the multiple linear regression equation to predict repair time, given the number of months since the last maintenance service and the type of repair. What are the interpretations of the estimated parameters? What does the coefficient of determination tell you about this model?

e. Develop the multiple linear regression equation to predict repair time, given the number of months since the last maintenance service, the type of repair, and the repairperson. What are the interpretations of the estimated parameters? What does the coefficient of determination tell you about this model?

f. Which of these models would you use? Why?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

14. **Delays in Company Audits.** A study investigated the relationship between audit delay (the length of time from a company's fiscal year-end to the date of the auditor's report) and variables that describe the client and the auditor. Some of the independent variables that were included in this study follow:

Industry	A dummy variable coded 1 if the firm was an industrial company or 0 if the firm was a bank, savings and loan, or insurance company.
Public	A dummy variable coded 1 if the company was traded on an organized exchange or over the counter; otherwise coded 0.
Quality	A measure of overall quality of internal controls, as judged by the auditor, on a 5-point scale ranging from "virtually none" (1) to "excellent" (5).
Finished	A measure ranging from 1 to 4, as judged by the auditor, where 1 indicates "all work performed subsequent to year-end" and 4 indicates "most work performed prior to year-end."

A sample of 40 companies provided the following data. **LO 2, 3, 4, 5**

DATA*file*
audit

Delay (Days)	Industry	Public	Quality	Finished
62	0	0	3	1
45	0	1	3	3
54	0	0	2	2
71	0	1	1	2
91	0	0	1	1
62	0	0	4	4
61	0	0	3	2
69	0	1	5	2
80	0	0	1	1
52	0	0	5	3
47	0	0	3	2
65	0	1	2	3
60	0	0	1	3
81	1	0	1	2
73	1	0	2	2
89	1	0	2	1
71	1	0	5	4
76	1	0	2	2
68	1	0	1	2
68	1	0	5	2
86	1	0	2	2
76	1	1	3	1
67	1	0	2	3
57	1	0	4	2
55	1	1	3	2
54	1	0	5	2
69	1	0	3	3
82	1	0	5	1
94	1	0	1	1
74	1	1	5	2

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Delay (Days)	Industry	Public	Quality	Finished
75	1	1	4	3
69	1	0	2	2
71	1	0	4	4
79	1	0	5	2
80	1	0	1	4
91	1	0	4	1
92	1	0	1	4
46	1	1	4	3
72	1	0	5	2
85	1	0	5	1

a. Develop the estimated linear regression equation using all of the independent variables included in the data.
b. How much of the variation in the sample values of delay does this estimated linear regression equation explain? What other independent variables could you include in this linear regression model to improve the fit?
c. Test the relationship between each independent variable and the dependent variable at the 0.05 level of significance, and interpret the relationship between each of the independent variables and the dependent variable.
d. On the basis of your observations about the relationships between the dependent variable Delay and the independent variables Quality and Finished, suggest an alternative model for the linear regression equation developed in part (a) to explain as much of the variability in Delay as possible using only the independent variables that were found to be statistically significant at the 0.05 level.

15. **Estimating Fuel Mileage by Car Size.** The U.S. Department of Energy's *Fuel Economy Guide* provides fuel efficiency data for cars and trucks. A portion of the data for 311 compact, midsized, and large cars follows. The Class column identifies the size of the car: Compact, Midsize, or Large. The Displacement column shows the engine's displacement in liters. The FuelType column shows whether the car uses premium (P) or regular (R) fuel, and the HwyMPG column shows the fuel efficiency rating for highway driving in terms of miles per gallon. The complete data set is contained in the file *fueldata*. **LO 1, 2, 3, 4, 5, 6, 7, 9**

DATA*file*
fueldata

Car	Class	Displacement	FuelType	HwyMPG
1	Compact	3.1	P	25
2	Compact	3.1	P	25
3	Compact	3.0	P	25
⋮	⋮	⋮	⋮	⋮
161	Midsize	2.4	R	30
162	Midsize	2.0	P	29
⋮	⋮	⋮	⋮	⋮
310	Large	3.0	R	25

a. Develop an estimated linear regression equation that can be used to predict the fuel efficiency for highway driving given the engine's displacement. Test for significance using the 0.05 level of significance. How much of the variation in the sample values of HwyMPG does this estimated linear regression equation explain?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Create a scatter chart with HwyMPG on the *y*-axis and displacement on the *x*-axis for which the points representing compact, midsize, and large automobiles are shown in different colors. What does this chart suggest about the relationship between the class of automobile (compact, midsize, and large) and HwyMPG?
c. Now consider the addition of the dummy variables ClassMidsize and ClassLarge to the simple linear regression model in part (a). The value of ClassMidsize is 1 if the car is a midsize car and 0 otherwise; the value of ClassLarge is 1 if the car is a large car and 0 otherwise. Thus, for a compact car, the value of ClassMidsize and the value of ClassLarge are both 0. Develop the estimated linear regression equation that can be used to predict the fuel efficiency for highway driving, given the engine's displacement and the dummy variables ClassMidsize and ClassLarge. How much of the variation in the sample values of HwyMPG is explained by this estimated linear regression equation?
d. Use significance level of 0.05 to determine whether the dummy variables added to the model in part (c) are significant.
e. Consider the addition of the dummy variable FuelPremium, where the value of FuelPremium is 1 if the car uses premium fuel and 0 if the car uses regular fuel. Develop the estimated linear regression equation that can be used to predict the fuel efficiency for highway driving given the engine's displacement, the dummy variables ClassMidsize and ClassLarge, and the dummy variable FuelPremium. How much of the variation in the sample values of HwyMPG does this estimated linear regression equation explain?
f. For the estimated linear regression equation developed in part (e), test for the significance of the relationship between each of the independent variables and the dependent variable using the 0.05 level of significance for each test.
g. An automobile manufacturer is designing a new compact model with a displacement of 2.9 liters with the objective of creating a model that will achieve at least 25 estimated highway MPG. The manufacturer must now decide if the car can be designed to use premium fuel and still achieve the objective of 25 MPG on the highway. Use the model developed in part (e) to recommend a decision to this manufacturer.

DATA*file* trafficflow

16. **Vehicle Speed and Traffic Flow.** A highway department is studying the relationship between traffic flow and speed during rush hour on Highway 193. The data in the file *trafficflow* were collected on Highway 193 during 100 recent rush hours. **LO 1, 2, 3, 4, 5, 8**
a. Develop a scatter chart for these data. What does the scatter chart indicate about the relationship between vehicle speed and traffic flow?
b. Develop an estimated simple linear regression equation for the data. How much variation in the sample values of traffic flow is explained by this linear regression model? Use a 0.05 level of significance to test the relationship between vehicle speed and traffic flow. What is the interpretation of this relationship?
c. Develop an estimated quadratic regression equation for the data. How much variation in the sample values of traffic flow is explained by this model? Test the relationship between each of the independent variables and the dependent variable at a 0.05 level of significance. How would you interpret this model? Is this model superior to the model you developed in part (b)?
d. As an alternative to fitting a second-order model, fit a model using a piecewise linear regression with a single knot. What value of vehicle speed appears to be a good point for the placement of the knot? Does the estimated piecewise linear regression model provide a better fit than the estimated quadratic regression model developed in part (c)? Explain.
e. Separate the data into two sets such that one data set contains the observations of vehicle speed less than the value of the knot from part (d) and the other data set contains the observations of vehicle speed greater than or equal to the value of the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

knot from part (d). Then fit a simple linear regression equation to each data set. How does this pair of linear regression equations compare to the single piecewise linear regression equation with the single knot from part (d)? In particular, compare predicted values of traffic flow for values of the speed slightly above and slightly below the knot value from part (d).

f. What other independent variables could you include in your regression model to explain more variation in traffic flow?

DATA*file* corporatebonds

17. **Years to Maturity and Bond Yield.** A sample containing years to maturity and (percent) yield for 40 corporate bonds is contained in the file *corporatebonds* (*Barron's*, April 2, 2012). **LO 1, 2, 3, 4, 5, 8**
    a. Develop a scatter chart of the data using years to maturity as the independent variable. Does a simple linear regression model appear to be appropriate?
    b. Develop an estimated quadratic regression equation with years to maturity and squared values of years to maturity as the independent variables. How much variation in the sample values of yield is explained by this model? Test the relationship between each of the independent variables and the dependent variable at a 0.05 level of significance. How would you interpret this model?
    c. Create a plot of the linear and quadratic regression models overlaid on the scatter chart of years to maturity and yield. Does this help you better understand the difference in how the quadratic regression model and a simple linear regression model fit the sample data? Which model does this chart suggest provides a superior fit to the sample data?
    d. What other independent variables could you include in your regression model to explain more variation in yield?

18. **Cost of Renting or Purchasing a Home.** In 2011, home prices and mortgage rates fell so far that in a number of cities the monthly cost of owning a home was less expensive than renting. The following data show the average asking rent for 10 markets and the monthly mortgage on the median priced home (including taxes and insurance) for 10 cities where the average monthly mortgage payment was less than the average asking rent (*The Wall Street Journal*, November 26–27, 2011). **LO 1, 2, 8**

DATA*file* rentmortgage

City	Rent ($)	Mortgage ($)
Atlanta	840	539
Chicago	1,062	1,002
Detroit	823	626
Jacksonville	779	711
Las Vegas	796	655
Miami	1,071	977
Minneapolis	953	776
Orlando	851	695
Phoenix	762	651
St. Louis	723	654

    a. Develop a scatter chart for these data, treating the average asking rent as the independent variable. Does a simple linear regression model appear to be appropriate?
    b. Use a simple linear regression model to develop an estimated linear regression equation to predict the monthly mortgage on the median-priced home given the average asking rent. Construct a plot of the residuals against the independent variable rent. Based on this residual plot, does a simple linear regression model appear to be appropriate?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Using a quadratic regression model, develop an estimated regression equation to predict the monthly mortgage on the median-priced home, given the average asking rent.
d. Do you prefer the estimated regression equation developed in part (a) or part (c)? Create a plot of the linear and quadratic regression models overlaid on the scatter chart of the monthly mortgage on the median-priced home and the average asking rent to help you assess the two regression equations. Explain your conclusions.

DATA*file* emphysema

19. **Factors in Emphysema Risk.** A recent 10-year study conducted by a research team at the Great Falls Medical School was conducted to assess how age, systolic blood pressure, and smoking relate to the risk of emphysema. Assume that the data in the file *emphysema* are from a portion of this study. Risk is interpreted as the probability (times 100) that the patient will develop emphysema over the next 10-year period. For the smoking variable, define a dummy variable with 1 indicating a smoker and 0 indicating a nonsmoker. **LO 2, 4, 5, 9**
a. Develop an estimated multiple linear regression equation that relates risk of developing emphysema to the person's age, systolic blood pressure, and whether the person is a smoker.
b. Is smoking a significant factor in the risk of developing emphysema? Explain. Use a 0.05 level of significance.
c. What is the probability of developing emphysema over the next 10 years for Art Speen, a 68-year-old smoker who has a systolic blood pressure of 175? What action might the physician recommend for this patient?
d. An insurance company will only sell its Select policy to people for whom the probability of developing emphysema in the next 10 years is less than 0.01. If a smoker with a systolic blood pressure of 180 applies for a Select policy, under what condition will the company sell the smoker the policy if it adheres to this standard?

DATA*file* rugglescollege

20. **GPA and SAT Scores.** The Scholastic Aptitude Test (or SAT) is a standardized college entrance test that is used by colleges and universities as a means for making admission decisions. The critical reading and mathematics components of the SAT are reported on a scale from 200 to 800. Several universities believe these scores are strong predictors of an incoming student's potential success, and they use these scores as inputs when making admission decisions on potential freshman. The file *rugglescollege* contains freshman year GPA and the critical reading and mathematics SAT scores for a random sample of 200 students who recently completed their freshman year at Ruggles College. **LO 2, 4, 5, 9**
a. Develop an estimated multiple linear regression equation that includes critical reading and mathematics SAT scores as independent variables. How much variation in freshman GPA is explained by this model? Test whether each of the parameters β_0, β_1, and β_2 is equal to zero at a 0.05 level of significance. What are the correct interpretations of the estimated parameters? Are these interpretations reasonable?
b. Using the multiple linear regression model you developed in part (a), what is the predicted freshman GPA of Bobby Engle, a student who has been admitted to Ruggles College with a 660 SAT score on critical reading and at a 630 SAT score on mathematics?
c. The Ruggles College Director of Admissions believes that the relationship between a student's scores on the critical reading component of the SAT and the student's freshman GPA varies with the student's score on the mathematics component of the SAT. Develop an estimated multiple linear regression equation that includes critical reading and mathematics SAT scores and their interaction as independent variables. How much variation in freshman GPA is explained by this model? Test whether each of the parameters β_0, β_1, β_2, and β_3 is equal to zero at a 0.05 level

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

of significance. What are the correct interpretations of the estimated parameters? Do these results support the conjecture made by the Ruggles College Director of Admissions?

d. Do you prefer the estimated linear regression model developed in part (a) or part (c)? Explain.

e. What other factors could be included in the model as independent variables?

DATA *file*
extendedlargecredit

21. **Consumer Credit Card Debt.** Consider again the example introduced in Section 8.9 of a credit card company that has a database of information provided by its customers when they apply for credit cards. An analyst has created a multiple regression model for which the dependent variable in the model is credit card charges accrued by a customer in the data set over the past year (y), and the independent variables are the customer's annual household income (x_1), number of members of the household (x_2), and number of years of post–high school education (x_3). Figure 8.37 provides Excel output for a multiple regression model estimated using a data set the company created. **LO 1, 2, 3, 7**

a. Estimate the corresponding simple linear regression with the customer's annual household income as the independent variable and credit card charges accrued by a customer over the past year as the dependent variable. Interpret the estimated relationship between the customer's annual household income and credit card charges accrued over the past year. How much variation in credit card charges accrued by a customer over the past year is explained by this simple linear regression model?

b. Estimate the corresponding simple linear regression with the number of members in the customer's household as the independent variable and credit card charges accrued by a customer over the past year as the dependent variable. Interpret the estimated relationship between the number of members in the customer's household and credit card charges accrued over the past year. How much variation in credit card charges accrued by a customer over the past year is explained by this simple linear regression model?

c. Estimate the corresponding simple linear regression with the customer's number of years of post–high school education as the independent variable and credit card charges accrued by a customer over the past year as the dependent variable. Interpret the estimated relationship between the customer's number of years of post–high school education and credit card charges accrued over the past year. How much variation in credit card charges accrued by a customer over the past year is explained by this simple linear regression model?

d. Recall the multiple linear regression in Figure 8.37 with credit card charges accrued by a customer over the past year as the dependent variable and customer's annual house–hold income (x_1), number of members of the household (x_2), and number of years of post–high school education (x_3) as the independent variables. Do the estimated slopes differ substantially from the corresponding slopes that were estimated using simple linear regression in parts (a), (b), and (c)? What does this tell you about multicollinearity in the multiple linear regression model in Figure 8.37?

e. Add the coefficients of determination for the simple linear regression in parts (a), (b), and (c), and compare the result to the coefficient of determination for the multiple linear regression model in Figure 8.37. What does this tell you about multicollinearity in the multiple linear regression model in Figure 8.37?

f. Add age, a dummy variable for sex, and a dummy variable for whether a customer has exceeded their credit limit in the past 12 months as independent variables to the multiple linear regression model in Figure 8.37. Code the dummy variable for sex as 1 if the customer is female and 0 if male, and code the dummy variable for whether a customer has exceeded their credit limit in the past 12 months as 1 if the customer has exceeded their credit limit in the past 12 months and 0 otherwise. Do these variables substantially improve the fit of your model?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Case Problem 1: Alumni Giving

Alumni donations are an important source of revenue for colleges and universities. If administrators could determine the factors that could lead to increases in the percentage of alumni who make a donation, they might be able to implement policies that could lead to increased revenues. Research shows that students who are more satisfied with their contact with teachers are more likely to graduate. As a result, one might suspect that smaller class sizes and lower student/faculty ratios might lead to a higher percentage of satisfied graduates, which in turn might lead to increases in the percentage of alumni who make a donation. The following table shows data for 48 national universities. The Graduation Rate column is the percentage of students who initially enrolled at the university and graduated. The % of Classes Under 20 column shows the percentages of classes with fewer than 20 students that are offered. The Student/Faculty Ratio column is the number of students enrolled divided by the total number of faculty. Finally, the Alumni Giving Rate column is the percentage of alumni who made a donation to the university. **LO 1, 2, 3, 4, 5, 9**

DATA*file*
alumnigiving

	State	Graduation Rate	% of Classes Under 20	Student/ Faculty Ratio	Alumni Giving Rate
Boston College	MA	85	39	13	25
Brandeis University	MA	79	68	8	33
Brown University	RI	93	60	8	40
California Institute of Technology	CA	85	65	3	46
Carnegie Mellon University	PA	75	67	10	28
Case Western Reserve Univ.	OH	72	52	8	31
College of William and Mary	VA	89	45	12	27
Columbia University	NY	90	69	7	31
Cornell University	NY	91	72	13	35
Dartmouth College	NH	94	61	10	53
Duke University	NC	92	68	8	45
Emory University	GA	84	65	7	37
Georgetown University	DC	91	54	10	29
Harvard University	MA	97	73	8	46
Johns Hopkins University	MD	89	64	9	27
Lehigh University	PA	81	55	11	40
Massachusetts Institute of Technology	MA	92	65	6	44
New York University	NY	72	63	13	13
Northwestern University	IL	90	66	8	30
Pennsylvania State Univ.	PA	80	32	19	21
Princeton University	NJ	95	68	5	67
Rice University	TX	92	62	8	40
Stanford University	CA	92	69	7	34
Tufts University	MA	87	67	9	29
Tulane University	LA	72	56	12	17
University of California–Berkeley	CA	83	58	17	18
University of California–Davis	CA	74	32	19	7

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

	State	Graduation Rate	% of Classes Under 20	Student/ Faculty Ratio	Alumni Giving Rate
University of California–Irvine	CA	74	42	20	9
University of California– Los Angeles	CA	78	41	18	13
University of California–San Diego	CA	80	48	19	8
University of California–Santa Barbara	CA	70	45	20	12
University of Chicago	IL	84	65	4	36
University of Florida	FL	67	31	23	19
University of Illinois–Urbana Champaign	IL	77	29	15	23
University of Michigan–Ann Arbor	MI	83	51	15	13
University of North Carolina–Chapel Hill	NC	82	40	16	26
University of Notre Dame	IN	94	53	13	49
University of Pennsylvania	PA	90	65	7	41
University of Rochester	NY	76	63	10	23
University of Southern California	CA	70	53	13	22
University of Texas–Austin	TX	66	39	21	13
University of Virginia	VA	92	44	13	28
University of Washington	WA	70	37	12	12
University of Wisconsin–Madison	WI	73	37	13	13
Vanderbilt University	TN	82	68	9	31
Wake Forest University	NC	82	59	11	38
Washington University–St. Louis	MO	86	73	7	33
Yale University	CT	94	77	7	50

Managerial Report

1. Use methods of descriptive statistics to summarize the data.
2. Develop an estimated simple linear regression model that can be used to predict the alumni giving rate, given the graduation rate. Discuss your findings.
3. Develop an estimated multiple linear regression model that could be used to predict the alumni giving rate using Graduation Rate, % of Classes Under 20, and Student/Faculty Ratio as independent variables. Discuss your findings.
4. Based on the results in parts (2) and (3), do you believe another model may be more appropriate? Estimate this model, and discuss your results.
5. What conclusions and recommendations can you derive from your analysis? What universities are achieving a substantially higher alumni giving rate than would be expected, given their Graduation Rate, % of Classes Under 20, and Student/Faculty Ratio? What universities are achieving a substantially lower alumni giving rate than would be expected, given their Graduation Rate, % of Classes Under 20, and Student/Faculty Ratio? What other independent variables could be included in the model?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Case Problem 2: Consumer Research, Inc.

Consumer Research, Inc., is an independent agency that conducts research on consumer attitudes and behaviors for a variety of firms. In one study, a client asked for an investigation of consumer characteristics that can be used to predict the amount charged by credit card users. Data were collected on annual income, household size, and annual credit card charges for a sample of 50 consumers. The following data are contained in the file *consumer*. **LO 1, 2, 3, 4, 5, 9**

DATA*file* consumer

Income ($1000s)	Household Size	Amount Charged ($)	Income ($1000s)	Household Size	Amount Charged ($)
54	3	4,016	54	6	5,573
30	2	3,159	30	1	2,583
32	4	5,100	48	2	3,866
50	5	4,742	34	5	3,586
31	2	1,864	67	4	5,037
55	2	4,070	50	2	3,605
37	1	2,731	67	5	5,345
40	2	3,348	55	6	5,370
66	4	4,764	52	2	3,890
51	3	4,110	62	3	4,705
25	3	4,208	64	2	4,157
48	4	4,219	22	3	3,579
27	1	2,477	29	4	3,890
33	2	2,514	39	2	2,972
65	3	4,214	35	1	3,121
63	4	4,965	39	4	4,183
42	6	4,412	54	3	3,730
21	2	2,448	23	6	4,127
44	1	2,995	27	2	2,921
37	5	4,171	26	7	4,603
62	6	5,678	61	2	4,273
21	3	3,623	30	2	3,067
55	7	5,301	22	4	3,074
42	2	3,020	46	5	4,820
41	7	4,828	66	4	5,149

Source: Consumer Research, Inc. (https://www.bbb.org/us/ny/rochester/profile/secret-shopper/consumer-research-inc-0041-45625697)

Managerial Report

1. Use methods of descriptive statistics to summarize the data. Comment on the findings.
2. Develop estimated linear regression equations, first using annual income as the independent variable and then using household size as the independent variable. Which variable is the better predictor of annual credit card charges? Discuss your findings.
3. Develop an estimated linear regression equation with annual income and household size as the independent variables. Discuss your findings.
4. What is the predicted annual credit card charge for a three-person household with an annual income of $40,000?
5. Discuss the need for other independent variables that could be added to the model. What additional variables might be helpful?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Case Problem 3: Predicting Winnings for NASCAR Drivers

Matt Kenseth won the 2012 Daytona 500, the most important race of the NASCAR season. Matt's win was no surprise because for the 2011 season Matt finished fourth in the point standings with 2,330 points, behind Tony Stewart (2,403 points), Carl Edwards (2,403 points), and Kevin Harvick (2,345 points). In 2011 Matt earned $6,183,580 by winning three Poles (fastest driver in qualifying), winning three races, finishing in the top five 12 times, and finishing in the top ten 20 times. NASCAR's point system in 2011 allocated 43 points to the driver who finished first, 42 points to the driver who finished second, and so on down to 1 point for the driver who finished in the 43rd position. In addition any driver who led a lap received 1 bonus point, the driver who led the most laps received an additional bonus point, and the race winner was awarded 3 bonus points. But, the maximum number of points a driver could earn in any race was 48. The following table shows data for the 2011 season for the top 35 drivers (NASCAR website). These data are contained in the file *nascar*. **LO 1, 2, 3, 4, 5**

Managerial Report

1. Suppose you wanted to predict Winnings ($) using only the number of poles won (Poles), the number of wins (Wins), the number of top five finishes (Top 5), or the number of top ten finishes (Top 10). Which of these four variables provides the best single predictor of winnings?
2. Develop an estimated linear regression equation that can be used to predict Winnings ($) given the number of poles won (Poles), the number of wins (Wins), the number of top five finishes (Top 5), and the number of top ten finishes (Top 10). Test for individual significance and discuss your findings and conclusions.
3. Create two new independent variables: Top 2–5 and Top 6–10. Top 2–5 represents the number of times the driver finished between second and fifth place and Top 6–10

DATA *file*
nascar

Driver	Points	Poles	Wins	Top 5	Top 10	Winnings ($)
Tony Stewart	2,403	1	5	9	19	6,529,870
Carl Edwards	2,403	3	1	19	26	8,485,990
Kevin Harvick	2,345	0	4	9	19	6,197,140
Matt Kenseth	2,330	3	3	12	20	6,183,580
Brad Keselowski	2,319	1	3	10	14	5,087,740
Jimmie Johnson	2,304	0	2	14	21	6,296,360
Dale Earnhardt Jr.	2,290	1	0	4	12	4,163,690
Jeff Gordon	2,287	1	3	13	18	5,912,830
Denny Hamlin	2,284	0	1	5	14	5,401,190
Ryan Newman	2,284	3	1	9	17	5,303,020
Kurt Busch	2,262	3	2	8	16	5,936,470
Kyle Busch	2,246	1	4	14	18	6,161,020
Clint Bowyer	1,047	0	1	4	16	5,633,950
Kasey Kahne	1,041	2	1	8	15	4,775,160
A. J. Allmendinger	1,013	0	0	1	10	4,825,560
Greg Biffle	997	3	0	3	10	4,318,050
Paul Menard	947	0	1	4	8	3,853,690
Martin Truex Jr.	937	1	0	3	12	3,955,560
Marcos Ambrose	936	0	1	5	12	4,750,390
Jeff Burton	935	0	0	2	5	3,807,780
Juan Montoya	932	2	0	2	8	5,020,780
Mark Martin	930	2	0	2	10	3,830,910
David Ragan	906	2	1	4	8	4,203,660

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Driver	Points	Poles	Wins	Top 5	Top 10	Winnings ($)
Joey Logano	902	2	0	4	6	3,856,010
Brian Vickers	846	0	0	3	7	4,301,880
Regan Smith	820	0	1	2	5	4,579,860
Jamie McMurray	795	1	0	2	4	4,794,770
David Reutimann	757	1	0	1	3	4,374,770
Bobby Labonte	670	0	0	1	2	4,505,650
David Gilliland	572	0	0	1	2	3,878,390
Casey Mears	541	0	0	0	0	2,838,320
Dave Blaney	508	0	0	1	1	3,229,210
Andy Lally	398	0	0	0	0	2,868,220
Robby Gordon	268	0	0	0	0	2,271,890
J. J. Yeley	192	0	0	0	0	2,559,500

Source: NASCAR website, February 28, 2011. (https://www.nascar.com/)

represents the number of times the driver finished between sixth and tenth place. Develop an estimated linear regression equation that can be used to predict Winnings ($) using Poles, Wins, Top 2–5, and Top 6–10. Test for individual significance and discuss your findings and conclusions.

4. Based upon the results of your analysis, which estimated linear regression equation would you recommend using to predict Winnings ($)? Provide an interpretation of the estimated coefficients for this equation.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 9

Time Series Analysis and Forecasting

Contents

Analytics in Action: ACCO Brands

9.1 Time Series Patterns
- Horizontal Pattern
- Trend Pattern
- Seasonal Pattern
- Trend and Seasonal Pattern
- Cyclical Pattern
- Identifying Time Series Patterns

9.2 Forecast Accuracy

9.3 Moving Averages and Exponential Smoothing
- Moving Averages
- Exponential Smoothing

9.4 Using Linear Regression Analysis for Forecasting
- Linear Trend Projection
- Seasonality Without Trend
- Seasonality with Trend
- Using Linear Regression Analysis as a Causal Forecasting Method
- Combining Causal Variables with Trend and Seasonality Effects
- Considerations in Using Linear Regression in Forecasting

9.5 Determining the Best Forecasting Model to Use

Summary 505
Glossary 505
Problems 506
Appendix 9.1: Using the Excel Forecast Sheet 517

Available in the Cengage eBook:
- Appendix: Forecasting with R

Learning Objectives

After completing this chapter, you will be able to

LO 1 Calculate measures of forecasting accuracy including mean absolute error, mean squared error, and mean absolute percentage error.

LO 2 Use measures of forecasting accuracy to choose an appropriate forecasting model.

LO 3 Identify an underlying pattern from a time-series plot of data.

LO 4 Use simple techniques such as the naïve method and the average of all historical data to forecast a time series that exhibits a horizontal pattern.

LO 5 Use smoothing techniques such as moving average and exponential smoothing to forecast a time series that exhibits a horizontal pattern.

LO 6 Use simple linear regression to forecast a time series that exhibits a linear trend.

LO 7 Use multiple linear regression analysis to develop a forecast for a time series.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

ACCO Brands*

ACCO Brands Corporation is one of the world's largest suppliers of branded office and consumer products and print finishing solutions. The company's brands include AT-A-GLANCE®, Day-Timer®, Five Star®, GBC®, Hilroy®, Kensington®, Marbig®, Mead®, NOBO, Quartet®, Rexel, Swingline®, Tilibra®, Wilson Jones®, and many others.

Because it produces and markets a wide array of products with myriad demand characteristics, ACCO Brands relies heavily on sales forecasts in planning its manufacturing, distribution, and marketing activities. By viewing its relationship in terms of a supply chain, ACCO Brands and its customers (which are generally retail chains) establish close collaborative relationships and consider each other to be valued partners. As a result, ACCO Brands' customers share valuable information and data that serve as inputs into ACCO Brands' forecasting process.

In her role as a forecasting manager for ACCO Brands, Vanessa Baker appreciates the importance of this additional information. "We do separate forecasts of demand for each major customer," said Baker, "and we generally use twenty-four to thirty-six months of history to generate monthly forecasts twelve to eighteen months into the future. While trends are important, several of our major product lines, including school, planning and organizing, and decorative calendars, are heavily seasonal, and seasonal sales make up the bulk of our annual volume."

Daniel Marks, one of several account-level strategic forecast managers for ACCO Brands, adds:

> The supply chain process includes the total lead time from identifying opportunities to making or procuring the product to getting the product on the shelves to align with the forecasted demand; this can potentially take several months, so the accuracy of our forecasts is critical throughout each step of the supply chain. Adding to this challenge is the risk of obsolescence. We sell many dated items, such as planners and calendars, which have a natural, built-in obsolescence. In addition, many of our products feature designs that are fashion-conscious or contain pop culture images, and these products can also become obsolete very quickly as tastes and popularity change. An overly optimistic forecast for these products can be very costly, but an overly pessimistic forecast can result in lost sales potential and give our competitors an opportunity to take market share from us.

In addition to trends, seasonal components, and cyclical patterns, there are several other factors that Baker and Marks must consider. Baker notes, "We have to adjust our forecasts for upcoming promotions by our customers." Marks agrees and adds:

> We also have to go beyond just forecasting consumer demand; we must consider the retailer's specific needs in our order forecasts, such as what type of display will be used and how many units of a product must be on display to satisfy their presentation requirements. Current inventory is another factor—if a customer is carrying either too much or too little inventory, that will affect their future orders, and we need to reflect that in our forecasts. Will the product have a short life because it is tied to a cultural fad? What are the retailer's marketing and markdown strategies? Our knowledge of the environments in which our supply chain partners are competing helps us to forecast demand more accurately, and that reduces waste and makes our customers, as well as ACCO Brands, far more profitable.

*The authors are indebted to Vanessa Baker and Daniel Marks of ACCO Brands for providing input for this Analytics in Action.

The purpose of this chapter is to provide an introduction to time series analysis and forecasting. Suppose we are asked to provide quarterly **forecasts** of sales for one of our company's products over the upcoming one-year period. Production schedules, raw materials purchasing, inventory policies, marketing plans, and cash flows will all be affected by the quarterly forecasts we provide. Consequently, poor forecasts may result in poor planning and increased costs for the company. How should we go about providing the quarterly sales forecasts? Good judgment, intuition, and an awareness of the state of the economy may give us a rough idea, or feeling, of what is likely to happen in the future, but converting that feeling into a number that can be used as next year's sales forecast is challenging.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A forecast is simply a prediction of what will happen in the future. Managers must accept that regardless of the technique used, they will not be able to develop perfect forecasts.

Forecasting methods can be classified as qualitative or quantitative. Qualitative methods generally involve the use of expert judgment to develop forecasts. Such methods are appropriate when historical data on the variable being forecast are either unavailable or not applicable. Quantitative forecasting methods can be used when (1) past information about the variable being forecast is available, (2) the information can be quantified, and (3) it is reasonable to assume that past is prologue (i.e., that the pattern of the past will continue into the future). In this chapter, we will focus exclusively on quantitative forecasting methods.

If the historical data are restricted to past values of the variable to be forecast, the forecasting procedure is called a time series method and the historical data are referred to as *time series*. The objective of time series analysis is to uncover a pattern in the time series and then extrapolate the pattern to forecast the future; the forecast is based solely on past values of the variable and/or on past forecast errors.

Causal or exploratory forecasting methods are based on the assumption that the variable we are forecasting has a cause-and-effect relationship with one or more other variables. These methods help explain how the value of one variable impacts the value of another. For instance, the sales volume for many products is influenced by advertising expenditures, so regression analysis may be used to develop an equation showing how these two variables are related. Then, once the advertising budget is set for the next period, we could substitute this value into the equation to develop a prediction or forecast of the sales volume for that period. Note that if a time series method was used to develop the forecast, advertising expenditures would not be considered; that is, a time series method would base the forecast solely on past sales.

Modern data-collection technologies have enabled individuals, businesses, and government agencies to collect vast amounts of data that may be used for causal forecasting. For example, supermarket scanners allow retailers to collect point-of-sale data that can then be used to help aid in planning sales, coupon targeting, and other marketing and planning efforts. These data can help answer important questions like, "Which products tend to be purchased together?" One of the techniques used to answer questions using such data is regression analysis. In this chapter, we discuss the use of regression analysis as a causal forecasting method.

In this chapter, we also discuss the various kinds of time series that a forecaster might be faced with in practice. These include a constant or horizontal pattern, a trend, a seasonal pattern, both a trend and a seasonal pattern, and a cyclical pattern. To build a quantitative forecasting model, it is also necessary to have a measurement of forecast accuracy. Different measurements of forecast accuracy, as well as their respective advantages and disadvantages, are discussed. For a horizontal or constant time series, we develop the classical moving average, weighted moving average, and exponential smoothing models. Many time series have a trend, and taking this trend into account is important; we provide regression models for finding the best model parameters when a linear trend is present, when the data show a seasonal pattern, or when the variable to be predicted has a causal relationship with other variables. Finally, we discuss considerations to be made when determining the best forecasting model to use. The chapter appendix provides an overview of using the Excel Forecast Sheet for time series forecasting.

Notes + Comments

Virtually all large companies today rely on enterprise resource planning (ERP) software to aid in their planning and operations. These software systems help the business run smoothly by collecting and efficiently storing company data, enabling it to be shared company-wide for planning at all levels: strategically, tactically, and operationally. Most ERP systems include a forecasting module to help plan for the future. SAP, one of the most widely used ERP systems, includes a forecasting component. This module allows the user to select from a number of forecasting techniques and/or have the system find a "best" model. The various forecasting methods and ways to measure the quality of a forecasting model discussed in this chapter are routinely available in software that supports forecasting.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

9.1 Time Series Patterns

We limit our discussion to time series for which the values of the series are recorded at equal intervals. Cases in which the observations are made at unequal intervals are beyond the scope of this text.

A **time series** is a sequence of observations on a variable measured at successive points in time or over successive periods of time. The measurements may be taken every hour, day, week, month, year, or at any other regular interval. The pattern of the data is an important factor in understanding how the time series has behaved in the past. If such behavior can be expected to continue in the future, we can use it to guide us in selecting an appropriate forecasting method.

To identify the underlying pattern in the data, a useful first step is to construct a *time series plot*, which is a graphical presentation of the relationship between time and the time series variable; time is represented on the horizontal axis and values of the time series variable are shown on the vertical axis. Let us first review some of the common types of data patterns that can be identified in a time series plot.

In Chapter 2 we discuss line charts, which are often used to graph time series.

Horizontal Pattern

A horizontal pattern exists when the data fluctuate randomly around a constant mean over time. To illustrate a time series with a horizontal pattern, consider the 12 weeks of data in Table 9.1. These data show the number of gallons of gasoline (in 1,000s) sold by a gasoline distributor in Bennington, Vermont, over the past 12 weeks. The average value, or mean, for this time series is 19.25 or 19,250 gallons per week. Figure 9.1 shows a time series plot for these data. Note how the data fluctuate around the sample mean of 19,250 gallons. Although random variability is present, we would say that these data follow a horizontal pattern.

The term **stationary time series** is used to denote a time series whose statistical properties are independent of time. In particular this means that:

1. The process generating the data has a constant mean.
2. The variability of the time series is constant over time.

For a formal definition of stationarity, see K. Ord,R. Fildes, and N. Kourentzes, Principles of Business Forecasting, *2nd ed. (Wessex, Inc., 2017).*

A time series plot for a stationary time series will always exhibit a horizontal pattern with random fluctuations. However, simply observing a horizontal pattern is not sufficient evidence to conclude that the time series is stationary. More advanced texts on forecasting discuss procedures for determining whether a time series is stationary and provide methods for transforming a nonstationary time series into a stationary series.

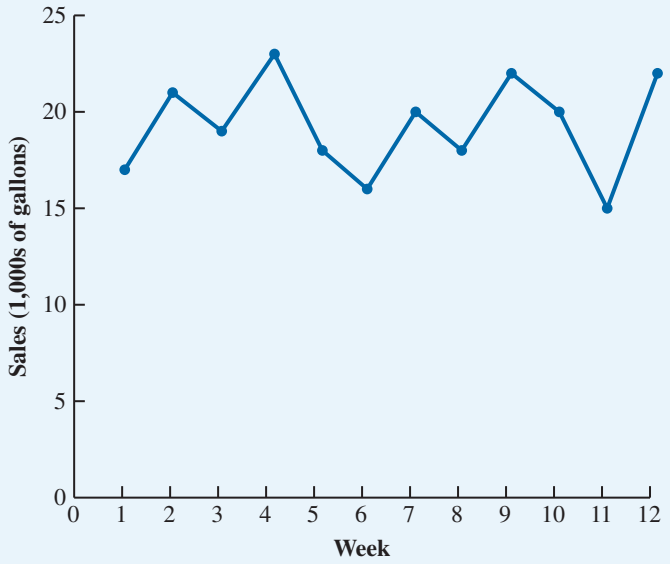
Table 9.1 Gasoline Sales Time Series

Week	Sales (1,000s of gallons)
1	17
2	21
3	19
4	23
5	18
6	16
7	20
8	18
9	22
10	20
11	15
12	22

DATA*file*
gasoline

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.1 Gasoline Sales Time Series Plot



Changes in business conditions often result in a time series with a horizontal pattern that shifts to a new level at some point in time. For instance, suppose the gasoline distributor signs a contract with the Vermont State Police to provide gasoline for state police cars located in southern Vermont beginning in week 13. With this new contract, the distributor naturally expects to see a substantial increase in weekly sales starting in week 13. Table 9.2 shows the number of gallons of gasoline sold for the original time series and for the 10 weeks after signing the new contract. Figure 9.2 shows the corresponding time series plot. Note the increased level of the time series beginning in week 13. This change in the

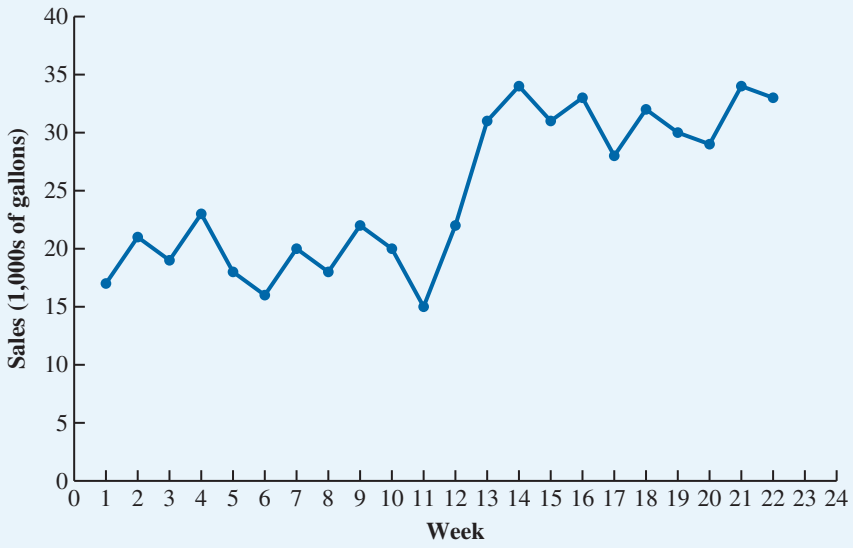
Table 9.2 Gasoline Sales Time Series After Obtaining the Contract with the Vermont State Police

Week	Sales (1,000s of gallons)	Week	Sales (1,000s of gallons)
1	17	12	22
2	21	13	31
3	19	14	34
4	23	15	31
5	18	16	33
6	16	17	28
7	20	18	32
8	18	19	30
9	22	20	29
10	20	21	34
11	15	22	33

DATA*file*
gasolinerevised

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.2 Gasoline Sales Time Series Plot After Obtaining the Contract with the Vermont State Police



level of the time series makes it more difficult to choose an appropriate forecasting method. Selecting a forecasting method that adapts well to changes in the level of a time series is an important consideration in many practical applications.

Trend Pattern

Although time series data generally exhibit random fluctuations, a time series may also show gradual shifts or movements to relatively higher or lower values over a longer period of time. If a time series plot exhibits this type of behavior, we say that a **trend** pattern exists. A trend is usually the result of long-term factors such as population increases or decreases, shifting demographic characteristics of the population, improving technology, changes in the competitive landscape, and/or changes in consumer preferences.

To illustrate a time series with a linear trend pattern, consider the time series of bicycle sales for a particular manufacturer over the past 10 years, as shown in Table 9.3 and Figure 9.3. Note that a total of 21,600 bicycles were sold in year 1, a total of 22,900 in year 2, and so on. In year 10, the most recent year, 31,400 bicycles were sold. Visual inspection of the time series plot shows some up-and-down movement over the past 10 years, but the time series seems also to have a systematically increasing, or upward, trend.

The trend for the bicycle sales time series appears to be linear and increasing over time, but sometimes a trend can be described better by other types of patterns. For instance, the data in Table 9.4 and the corresponding time series plot in Figure 9.4 show the sales revenue for a cholesterol drug since the company won FDA approval for the drug 10 years ago. The time series increases in a nonlinear fashion; that is, the rate of change of revenue does not increase by a constant amount from one year to the next. In fact, the revenue appears to be growing in an exponential fashion. Exponential relationships such as this are appropriate when the *percentage* change from one period to the next is relatively constant.

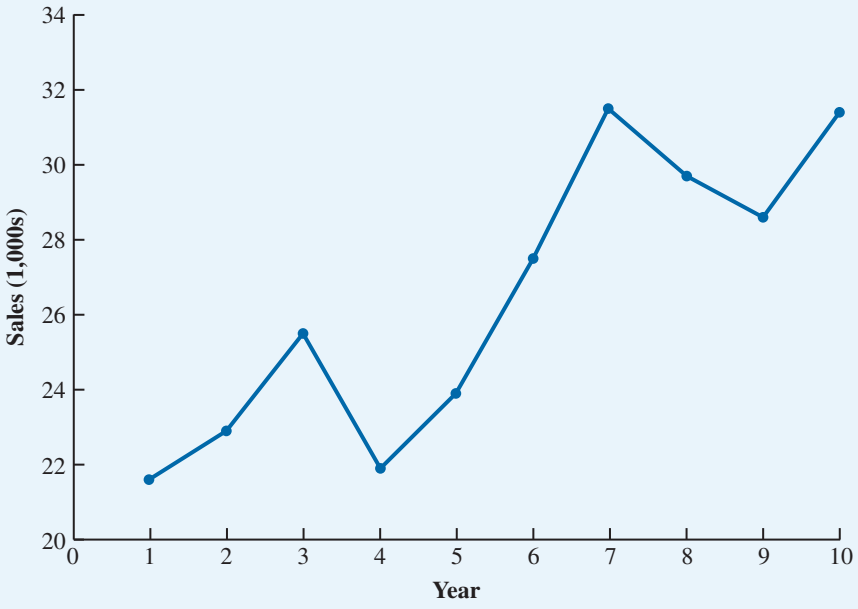
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
bicycle

Table 9.3 Bicycle Sales Time Series

Year	Sales (1,000s)
1	21.6
2	22.9
3	25.5
4	21.9
5	23.9
6	27.5
7	31.5
8	29.7
9	28.6
10	31.4

Figure 9.3 Bicycle Sales Time Series Plot



Seasonal Pattern

The trend of a time series can be identified by analyzing movements in historical data over multiple time periods. **Seasonal patterns** are recognized by observing recurring patterns over successive periods of time. For example, a retailer who sells bathing suits expects low sales activity in the fall and winter months, with peak sales in the spring and summer months to occur every year. Retailers who sell snow removal equipment and heavy clothing, however, expect the opposite yearly pattern. Not surprisingly, the pattern for a time series plot that exhibits a recurring pattern over a one-year period due to seasonal influences is called a seasonal pattern. Although we generally think of seasonal movement in a time series as occurring over one year, time series data can also exhibit seasonal patterns of less than one year in duration. For example, daily traffic volume shows within-the-day

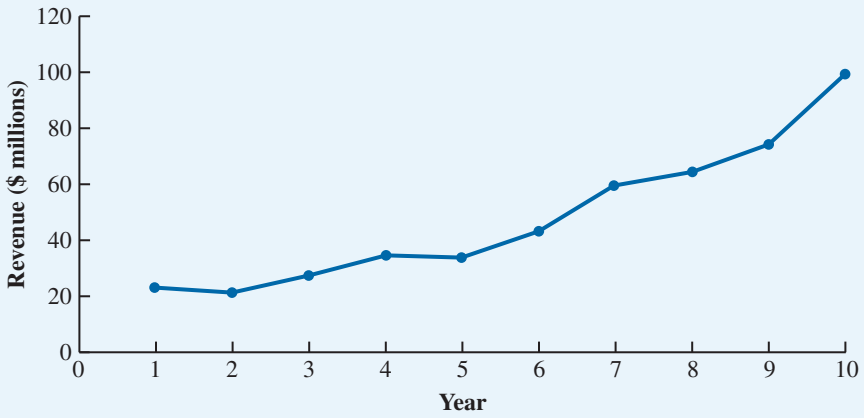
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
cholesterol

Table 9.4 Cholesterol Drug Revenue Time Series

Year	Revenue ($ millions)
1	23.1
2	21.3
3	27.4
4	34.6
5	33.8
6	43.2
7	59.5
8	64.4
9	74.2
10	99.3

Figure 9.4 Cholesterol Drug Revenue Time Series Plot ($ millions)



"seasonal" behavior, with peak levels occurring during rush hours, moderate flow during the rest of the day and early evening, and light flow from midnight to early morning. Another example of an industry with sales that exhibit easily discernible seasonal patterns within a day is the restaurant industry.

As an example of a seasonal pattern, consider the number of umbrellas sold at a clothing store over the past five years. Table 9.5 shows the time series and Figure 9.5 shows the corresponding time series plot. The time series plot does not indicate a long-term trend in sales. In fact, unless you look carefully at the data, you might conclude that the data follow a horizontal pattern with random fluctuation. However, closer inspection of the fluctuations in the time series plot reveals a systematic pattern in the data that occurs within each year. Specifically, the first and third quarters have moderate sales, the second quarter has the highest sales, and the fourth quarter has the lowest sales volume. Thus, we would conclude that a quarterly seasonal pattern is present.

Trend and Seasonal Pattern

Some time series include both a trend and a seasonal pattern. For instance, the data in Table 9.6 and the corresponding time series plot in Figure 9.6 show quarterly smartphone sales for a particular manufacturer over the past four years. Clearly an increasing trend is

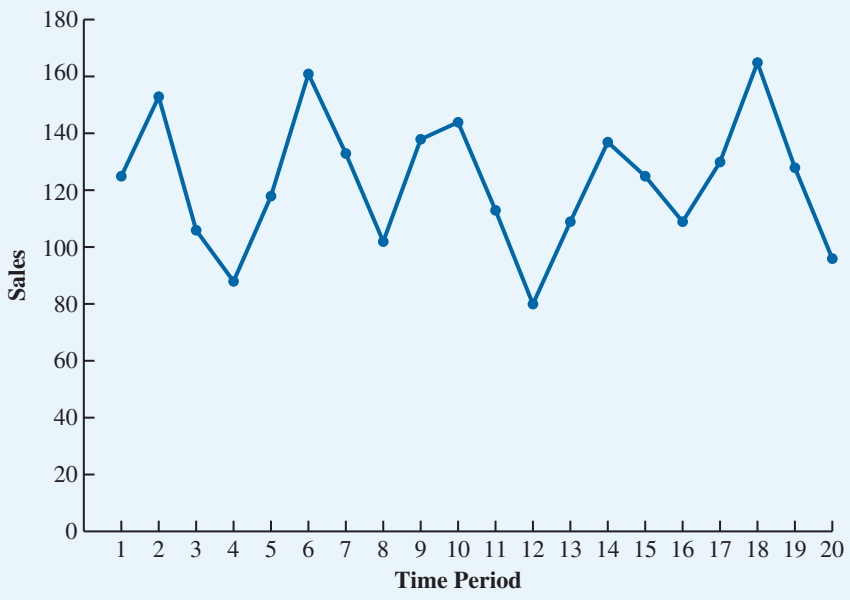
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
umbrella

Table 9.5 Umbrella Sales Time Series

Year	Quarter	Sales
1	1	125
	2	153
	3	106
	4	88
2	1	118
	2	161
	3	133
	4	102
3	1	138
	2	144
	3	113
	4	80
4	1	109
	2	137
	3	125
	4	109
5	1	130
	2	165
	3	128
	4	96

Figure 9.5 Umbrella Sales Time Series Plot



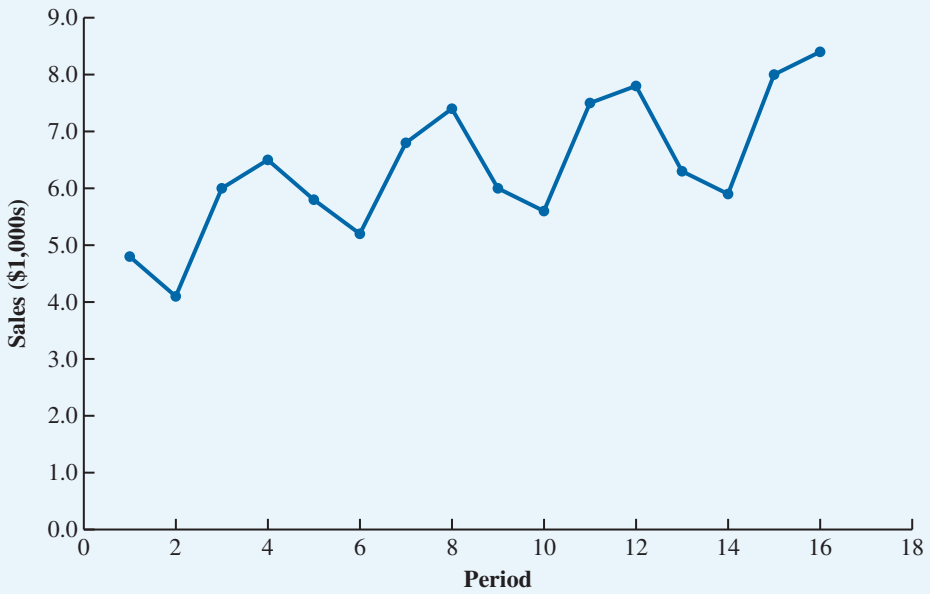
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
smartphonesales

Table 9.6 Quarterly Smartphone Sales Time Series

Year	Quarter	Sales ($1,000s)
1	1	4.8
	2	4.1
	3	6.0
	4	6.5
2	1	5.8
	2	5.2
	3	6.8
	4	7.4
3	1	6.0
	2	5.6
	3	7.5
	4	7.8
4	1	6.3
	2	5.9
	3	8.0
	4	8.4

Figure 9.6 Quarterly Smartphone Sales Time Series Plot



present. However, Figure 9.6 also indicates that sales are lowest in the second quarter of each year and highest in quarters 3 and 4. Thus, we conclude that a seasonal pattern also exists for smartphone sales. In such cases, we need to use a forecasting method that is capable of dealing with both trend and seasonality.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Cyclical Pattern

A **cyclical pattern** exists if the time series plot shows an alternating sequence of points below and above the trendline that lasts for more than one year. Many economic time series exhibit cyclical behavior with regular runs of observations below and above the trendline. Often the cyclical component of a time series is due to multiyear business cycles. For example, periods of moderate inflation followed by periods of rapid inflation can lead to a time series that alternates below and above a generally increasing trendline (e.g., a time series of housing costs). Business cycles are extremely difficult, if not impossible, to forecast. As a result, cyclical effects are often combined with long-term trend effects and referred to as *trend-cycle effects*. In this chapter, we do not deal with cyclical effects that may be present in the time series.

Identifying Time Series Patterns

The underlying pattern in the time series is an important factor in selecting a forecasting method. Thus, a time series plot should be one of the first analytic tools employed when trying to determine which forecasting method to use. If we see a horizontal pattern, then we need to select a method appropriate for this type of pattern. Similarly, if we observe a trend in the data, then we need to use a forecasting method that is capable of handling the conjectured type of trend effectively. In the next section, we discuss methods for assessing forecast accuracy.

9.2 Forecast Accuracy

In this section, we begin by developing forecasts for the gasoline time series shown in Table 9.1 using the simplest of all the forecasting methods. We use the most recent week's sales volume as the forecast for the next week. For instance, the distributor sold 17 thousand gallons of gasoline in week 1; this value is used as the forecast for week 2. Next, we use 21, the actual value of sales in week 2, as the forecast for week 3, and so on. The forecasts obtained for the historical data using this method are shown in Table 9.7 in the Forecast column. Because of its simplicity, this method is often referred to as a **naïve forecasting method**.

Table 9.7 Computing Forecasts and Measures of Forecast Accuracy Using the Most Recent Value as the Forecast for the Next Period

Week	Time Series Value	Forecast	Forecast Error	Absolute Value of Forecast Error	Squared Forecast Error	Percentage Error	Absolute Value of Percentage Error
1	17						
2	21	17	4	4	16	19.05	19.05
3	19	21	−2	2	4	−10.53	10.53
4	23	19	4	4	16	17.39	17.39
5	18	23	−5	5	25	−27.78	27.78
6	16	18	−2	2	4	−12.50	12.50
7	20	16	4	4	16	20.00	20.00
8	18	20	−2	2	4	−11.11	11.11
9	22	18	4	4	16	18.18	18.18
10	20	22	−2	2	4	−10.00	10.00
11	15	20	−5	5	25	−33.33	33.33
12	22	15	7	7	49	31.82	31.82
		Totals	5	41	179	1.19	211.69

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

How accurate are the forecasts obtained using this naïve forecasting method? To answer this question, we will introduce several measures of forecast accuracy. These measures are used to determine how well a particular forecasting method is able to reproduce the time series data that are already available. By selecting the method that is most accurate for the data already observed, we hope to increase the likelihood that we will obtain more accurate forecasts for future time periods. The key concept associated with measuring forecast accuracy is **forecast error**. If we denote y_t and $\hat{y}_t$ as the actual and forecasted values of the time series for period t, respectively, the forecasting error for period t is as follows:

Forecast Error

$$e_t = y_t - \hat{y}_t \tag{9.1}$$

That is, the forecast error for time period t is the difference between the actual and the forecasted values for period t.

For instance, because the distributor actually sold 21 thousand gallons of gasoline in week 2, and the forecast, using the sales volume in week 1, was 17 thousand gallons, the forecast error in week 2 is

$$e_2 = y_2 - \hat{y}_2 = 21 - 17 = 4$$

A positive error such as this indicates that the forecasting method underestimated the actual value of sales for the associated period. Next we use 21, the actual value of sales in week 2, as the forecast for week 3. Since the actual value of sales in week 3 is 19, the forecast error for week 3 is $e_3 = 19 - 21 = -2$. In this case, the negative forecast error indicates that the forecast overestimated the actual value for week 3. Thus, the forecast error may be positive or negative, depending on whether the forecast is too low or too high. A complete summary of the forecast errors for this naïve forecasting method is shown in Table 9.7 in the Forecast Error column. It is important to note that because we are using a past value of the time series to produce a forecast for period t, we do not have sufficient data to produce a naïve forecast for the first week of this time series.

A simple measure of forecast accuracy is the mean or average of the forecast errors. If we have n periods in our time series and k is the number of periods at the beginning of the time series for which we cannot produce a naïve forecast, the mean forecast error (MFE) is as follows:

Mean Forecast Error (MFE)

$$\text{MFE} = \frac{\sum_{t=k+1}^{n} e_t}{n-k} \tag{9.2}$$

Table 9.7 shows that the sum of the forecast errors for the gasoline sales time series is 5; thus, the mean, or average, error is $5/11 = 0.45$. Because we do not have sufficient data to produce a naïve forecast for the first week of this time series, we must adjust our calculations in both the numerator and denominator accordingly. This is common in forecasting; we often use k past periods from the time series to produce forecasts, and so we frequently cannot produce forecasts for the first k periods. In those instances, the summation in the numerator starts at the first value of t for which we have produced a forecast (so we begin the summation at $t = k + 1$), and the denominator (which is the number of periods in our time series for which we are able to produce a forecast) will also reflect these circumstances. In the gasoline example, although the time series consists of $n = 12$ values, to compute the mean error we divided the sum of the forecast errors by 11 because there are only 11 forecast errors (we cannot generate forecast sales for the first week using this naïve forecasting method).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Also note that in the gasoline time series, the mean forecast error is positive, implying that the method is generally under-forecasting; in other words, the observed values tend to be greater than the forecasted values. Because positive and negative forecast errors tend to offset each other, the mean error is likely to be small; thus, the mean error is not a very useful measure of forecast accuracy.

The **mean absolute error (MAE)** is a measure of forecast accuracy that avoids the problem of positive and negative forecast errors offsetting each other. As you might expect given its name, MAE is the average of the absolute values of the forecast errors:

The MAE is also referred to as the mean absolute deviation (MAD).

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{\sum_{t=k+1}^{n} |e_t|}{n-k} \tag{9.3}$$

Table 9.7 shows that the sum of the absolute values of the forecast errors is 41; thus:

$$\text{MAE} = \text{average of the absolute value of the forecast errors} = \frac{41}{11} = 3.73.$$

Another measure that avoids the problem of positive and negative errors offsetting each other is obtained by computing the average of the squared forecast errors. This measure of forecast accuracy is referred to as the **mean squared error (MSE)**:

Mean Squared Error (MSE)

$$\text{MSE} = \frac{\sum_{t=k+1}^{n} e_t^2}{n-k} \tag{9.4}$$

From Table 9.7, the sum of the squared errors is 179; hence

$$\text{MSE} = \text{average of the square of the forecast errors} = \frac{179}{11} = 16.27.$$

The size of the MAE or MSE depends on the scale of the data. As a result, it is difficult to make comparisons for different time intervals (such as comparing a method of forecasting monthly gasoline sales to a method of forecasting weekly sales) or to make comparisons across different time series (such as monthly sales of gasoline and monthly sales of oil filters). To make comparisons such as these we need to work with relative or percentage error measures. The **mean absolute percentage error (MAPE)** is such a measure. To calculate MAPE we use the following formula:

Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{\sum_{t=k+1}^{n} \left| \left(\frac{e_t}{y_t} \right) 100 \right|}{n-k} \tag{9.5}$$

Table 9.7 shows that the sum of the absolute values of the percentage errors is

$$\sum_{t=k+1}^{12} \left| \left(\frac{e_t}{y_t} \right) 100 \right| = 211.69$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Thus, the MAPE, which is the average of the absolute value of percentage forecast errors, is

$$\frac{211.69}{11} = 19.24\%$$

These measures of forecast accuracy simply measure how well the forecasting method is able to forecast historical values of the time series. Now, suppose we want to forecast sales for a future time period, such as week 13. The forecast for week 13 is 22, the actual value of the time series in week 12. Is this an accurate estimate of sales for week 13? Unfortunately, there is no way to address the issue of accuracy associated with forecasts for future time periods. However, if we select a forecasting method that works well for the historical data, and we have reason to believe the historical pattern will continue into the future, we should obtain forecasts that will ultimately be shown to be accurate.

Before concluding this section, let us consider another method for forecasting the gasoline sales time series in Table 9.1. Suppose we use the average of all the historical data available as the forecast for the next period. We begin by developing a forecast for week 2. Because there is only one historical value available prior to week 2, the forecast for week 2 is just the time series value in week 1; thus, the forecast for week 2 is 17 thousand gallons of gasoline. To compute the forecast for week 3, we take the average of the sales values in weeks 1 and 2. Thus

$$\hat{y}_3 = \frac{17 + 21}{2} = 19$$

Similarly, the forecast for week 4 is

$$\hat{y}_4 = \frac{17 + 21 + 19}{3} = 19$$

Table 9.8 Computing Forecasts and Measures of Forecast Accuracy Using the Average of All the Historical Data as the Forecast for the Next Period

Week	Time Series Value	Forecast	Forecast Error	Absolute Value of Forecast Error	Squared Forecast Error	Percentage Error	Absolute Value of Percentage Error
1	17						
2	21	17.00	4.00	4.00	16.00	19.05	19.05
3	19	19.00	0.00	0.00	0.00	0.00	0.00
4	23	19.00	4.00	4.00	16.00	17.39	17.39
5	18	20.00	−2.00	2.00	4.00	−11.11	11.11
6	16	19.60	−3.60	3.60	12.96	−22.50	22.50
7	20	19.00	1.00	1.00	1.00	5.00	5.00
8	18	19.14	−1.14	1.14	1.31	−6.35	6.35
9	22	19.00	3.00	3.00	9.00	13.64	13.64
10	20	19.33	0.67	0.67	0.44	3.33	3.33
11	15	19.40	−4.40	4.40	19.36	−29.33	29.33
12	22	19.00	3.00	3.00	9.00	13.64	13.64
		Totals	4.52	26.81	89.07	2.75	141.34

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The forecasts obtained using this method for the gasoline time series are shown in Table 9.8 in the Forecast column. Using the results shown in Table 9.8, we obtain the following values of MAE, MSE, and MAPE:

$$\text{MAE} = \frac{26.81}{11} = 2.44$$

$$\text{MSE} = \frac{89.07}{11} = 8.10$$

$$\text{MAPE} = \frac{141.34}{11} = 12.85\%$$

We can now compare the accuracy of the two forecasting methods we have considered in this section by comparing the values of MAE, MSE, and MAPE for each method.

	Naïve Method	Average of All Past Values
MAE	3.73	2.44
MSE	16.27	8.10
MAPE	19.24%	12.85%

As measured by MAE, MSE, and MAPE, the average of all past weekly gasoline sales provides more accurate forecasts for the next week than using the most recent week's gasoline sales.

Evaluating different forecasts based on historical accuracy is helpful only if historical patterns continue into the future. As we noted in Section 9.1, the 12 observations of Table 9.1 comprise a stationary time series. In Section 9.1, we also mentioned that changes in business conditions often result in a time series that is not stationary. We discussed a situation in which the gasoline distributor signed a contract with the Vermont State Police to provide gasoline for state police cars located in southern Vermont. Table 9.2 shows the number of gallons of gasoline sold for the original time series and for the 10 weeks after signing the new contract, and Figure 9.2 shows the corresponding time series plot. Note the change in level in week 13 for the resulting time series. When a shift to a new level such as this occurs, it takes several periods for the forecasting method that uses the average of all the historical data to adjust to the new level of the time series. However, in this case the simple naïve method adjusts very rapidly to the change in level because it uses only the most recent observation as the forecast.

Measures of forecast accuracy are important factors in comparing different forecasting methods, but we have to be careful not to rely too heavily on them. Good judgment and knowledge about business conditions that might affect the value of the variable to be forecast also have to be considered carefully when selecting a method. Historical forecast accuracy is not the sole consideration, especially if the pattern exhibited by the time series is likely to change in the future.

In the next section, we will introduce more sophisticated methods for developing forecasts for a time series that exhibits a horizontal pattern. Using the measures of forecast accuracy developed here, we will be able to assess whether such methods provide more accurate forecasts than we obtained using the simple approaches illustrated in this section. The methods that we will introduce also have the advantage that they adapt well to situations in which the time series changes to a new level. The ability of a forecasting method to adapt quickly to changes in level is an important consideration, especially in short-term forecasting situations.

9.3 Moving Averages and Exponential Smoothing

In this section, we discuss two forecasting methods that are appropriate for a time series with a horizontal pattern: moving averages and exponential smoothing. These methods are capable of adapting well to changes in the level of a horizontal pattern such as the one we saw with the extended gasoline sales time series (Table 9.2 and Figure 9.2). However,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

without modification they are not appropriate when considerable trend, cyclical, or seasonal effects are present. Because the objective of each of these methods is to smooth out random fluctuations in the time series, they are referred to as *smoothing methods*. These methods are easy to use and generally provide a high level of accuracy for short-range forecasts, such as a forecast for the next time period.

Moving Averages

The **moving average method** uses the average of the most recent k data values in the time series as the forecast for the next period. Mathematically, a moving average forecast of order k is:

Moving Average Forecast

$$\hat{y}_{t+1} = \frac{\Sigma(\text{most recent } k \text{ data values})}{k} = \frac{\sum_{i=t-k+1}^{t} y_i}{k} = \frac{y_{t-k+1} + \cdots + y_{t-1} + y_t}{k} \tag{9.6}$$

where

$\hat{y}_{t+1}$ = forecast of the time series for period $t + 1$
y_t = actual value of the time series in period t
k = number of periods of time series data used to generate the forecast

The term *moving* is used because every time a new observation becomes available for the time series, it replaces the oldest observation in the equation and a new average is computed. Thus, the periods over which the average is calculated change, or move, with each ensuing period.

To illustrate the moving averages method, let us return to the original 12 weeks of gasoline sales data in Table 9.1. The time series plot in Figure 9.1 indicates that the gasoline sales time series has a horizontal pattern. Thus, the smoothing methods of this section are applicable.

To use moving averages to forecast a time series, we must first select the order k, or the number of time series values to be included in the moving average. If only the most recent values of the time series are considered relevant, a small value of k is preferred. If a greater number of past values are considered relevant, then we generally opt for a larger value of k. As previously mentioned, a time series with a horizontal pattern can shift to a new level over time. A moving average will adapt to the new level of the series and continue to provide good forecasts in k periods. Thus, a smaller value of k will track shifts in a time series more quickly (the naïve approach discussed earlier is actually a moving average for $k = 1$). On the other hand, larger values of k will be more effective in smoothing out random fluctuations. Therefore, managerial judgment based on an understanding of the behavior of a time series is helpful in choosing an appropriate value of k.

To illustrate how moving averages can be used to forecast gasoline sales, we will use a three-week moving average ($k = 3$). We begin by computing the forecast of sales in week 4 using the average of the time series values in weeks 1 to 3:

$$\hat{y}_4 = \text{average for weeks 1 to 3} = \frac{17 + 21 + 19}{3} = 19$$

Thus, the moving average forecast of sales in week 4 is 19, or 19,000 gallons of gasoline. Because the actual value observed in week 4 is 23, the forecast error in week 4 is $e_4 = 23 - 19 = 4$.

We next compute the forecast of sales in week 5 by averaging the time series values in weeks 2 to 4:

$$\hat{y}_5 = \text{average for weeks 2 to 4} = \frac{21 + 19 + 23}{3} = 21$$

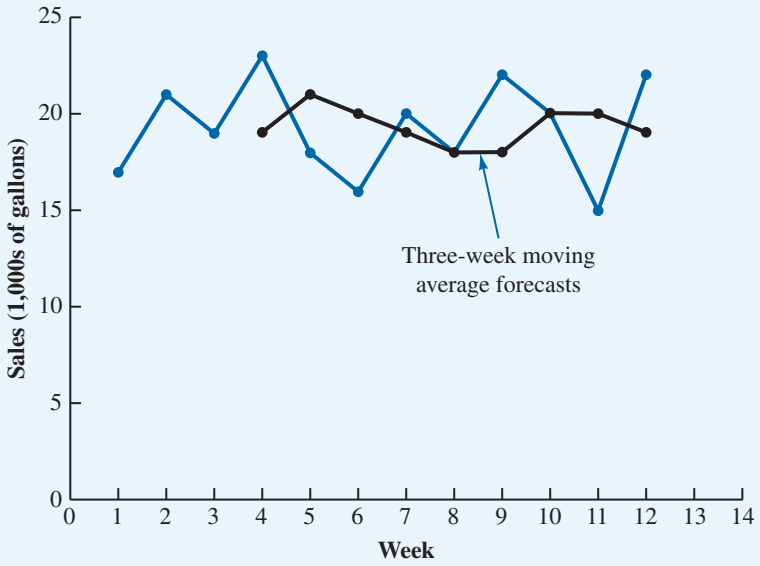
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Hence, the forecast of sales in week 5 is 21 and the error associated with this forecast is $e_5 = 18 - 21 = -3$. A complete summary of the three-week moving average forecasts for the gasoline sales time series is provided in Table 9.9. Figure 9.7 shows the original time series plot and the three-week moving average forecasts. Note how the graph of the moving average forecasts has tended to smooth out the random fluctuations in the time series.

Table 9.9 Summary of Three-Week Moving Average Calculations

Week	Time Series Value	Forecast	Forecast Error	Absolute Value of Forecast Error	Squared Forecast Error	Percentage Error	Absolute Value of Percentage Error
1	17						
2	21						
3	19						
4	23	19	4	4	16	17.39	17.39
5	18	21	−3	3	9	−16.67	16.67
6	16	20	−4	4	16	−25.00	25.00
7	20	19	1	1	1	5.00	5.00
8	18	18	0	0	0	0.00	0.00
9	22	18	4	4	16	18.18	18.18
10	20	20	0	0	0	0.00	0.00
11	15	20	−5	5	25	−33.33	33.33
12	22	19	3	3	9	13.64	13.64
		Totals	0	24	92	−20.79	129.21

Figure 9.7 Gasoline Sales Time Series Plot and Three-Week Moving Average Forecasts



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
gasoline

If **Data Analysis** *does not appear in the* **Analyze** *group in the* **Data** *tab, you will have to load the Analysis Toolpak add-in into Excel. To do so, click the* **File** *tab in the Ribbon and click* **Options**. *When the* **Excel Options** *dialog box appears, click* **Add-Ins** *from the menu. Next to* **Manage**, *select* **Excel Add-ins** *and click* **Go...** *at the bottom of the dialog box. When the* **Add-Ins** *dialog box appears, select* **Analysis Toolpak** *and click* **OK**.

To forecast sales in week 13, the next time period in the future, we simply compute the average of the time series values in weeks 10, 11, and 12:

$$\hat{y}_{13} = \text{average for weeks 10 to 12} = \frac{20 + 15 + 22}{3} = 19$$

Thus, the forecast for week 13 is 19, or 19,000 gallons of gasoline.

To show how Excel can be used to develop forecasts using the moving averages method, we develop a forecast for the gasoline sales time series in Table 9.1 and in the file *gasoline* as displayed in Columns A and B of Figure 9.10.

The following steps can be used to produce a three-week moving average:

Step 1. Click the **Data** tab in the Ribbon
Step 2. Click **Data Analysis** in the **Analyze** group
Step 3. When the **Data Analysis** dialog box appears (Figure 9.8), select **Moving Average** and click **OK**
Step 4. When the **Moving Average** dialog box appears (Figure 9.9):
Enter *B2:B13* in the **Input Range:** box
Enter *3* in the **Interval:** box
Enter *C3* in the **Output Range:** box
Click **OK**

Figure 9.8 Data Analysis Dialog Box

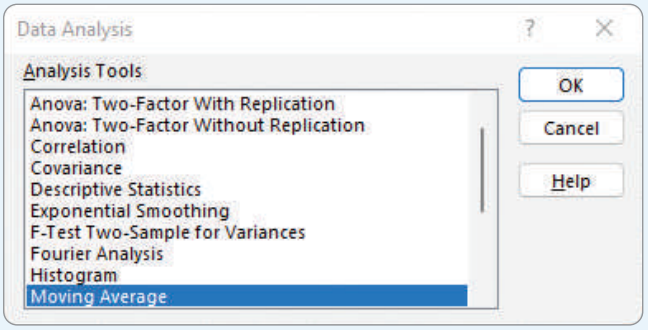
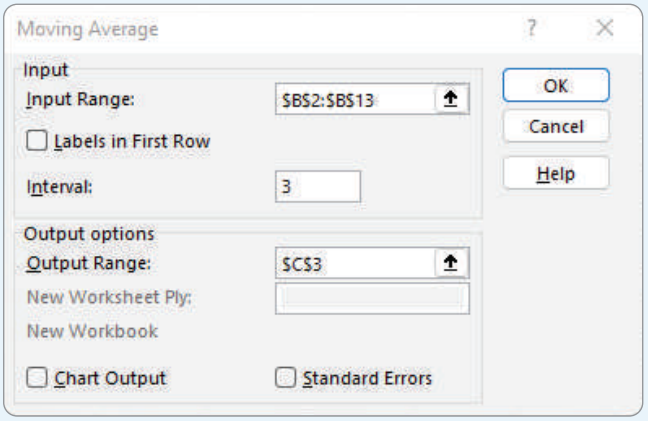


Figure 9.9 Moving Average Dialog Box



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.10 Excel Output for Moving Average Forecast for Gasoline Data

	A	B	C
1	Week	Sales (1000s of gallons)	
2	1	17	
3	2	21	#N/A
4	3	19	#N/A
5	4	23	19
6	5	18	21
7	6	16	20
8	7	20	19
9	8	18	18
10	9	22	18
11	10	20	20
12	11	15	20
13	12	22	19
14	13		19

Once you have completed this step, the three-week moving average forecasts will appear in column C of the worksheet as shown in Figure 9.10. Note that forecasts for periods of other lengths can be computed easily by entering a different value in the **Interval:** box.

Previously, we discussed three measures of forecast accuracy: mean absolute error (MAE), mean squared error (MSE), and mean absolute percentage error (MAPE). Using the three-week moving average calculations in Table 9.9, the values for these three measures of forecast accuracy are as follows:

$$\text{MAE} = \frac{\sum_{t=4}^{12} |e_t|}{n-3} = \frac{24}{9} = 2.67$$

$$\text{MSE} = \frac{\sum_{t=4}^{12} e_t^2}{n-3} = \frac{92}{9} = 10.22$$

$$\text{MAPE} = \frac{\sum_{t=4}^{12} \left| \left(\frac{e_t}{y_t} \right) 100 \right|}{n-3} = \frac{129.21}{9} = 14.36\%$$

In Section 9.2, we showed that using the most recent observation as the forecast for the next week (a moving average of order $k = 1$) resulted in values of MAE = 3.73, MSE = 16.27, and MAPE = 19.24%. Thus, according to each of these three measures, the three-week moving average approach has provided more accurate forecasts than simply using the most recent observation as the forecast. Also note how we have revised the formulas for the MAE, MSE, and MAPE to reflect that our use of a three-week moving average leaves us with insufficient data to generate forecasts for the first three weeks of our time series.

To determine whether a moving average with a different order k can provide more accurate forecasts, we recommend using trial and error to determine the value of k that minimizes the MSE. For the gasoline sales time series, it can be shown that the minimum

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

value of MSE corresponds to a moving average of order $k = 6$ with MSE = 6.79. If we are willing to assume that the order of the moving average that is best for the historical data will also be best for future values of the time series, the most accurate moving average forecasts of gasoline sales can be obtained using a moving average of order $k = 6$.

Exponential Smoothing

Exponential smoothing uses a weighted average of past time series values as a forecast. The exponential smoothing model is as follows:

Exponential Smoothing Forecast

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t \tag{9.7}$$

where

$\hat{y}_{t+1}$ = forecast of the time series for period $t + 1$
y_t = actual value of the time series in period t
$\hat{y}_t$ = forecast of the time series for period t
α = smoothing constant ($0 \leq \alpha \leq 1$)

Equation (9.7) shows that the forecast for period $t + 1$ is a weighted average of the actual value in period t and the forecast for period t. The weight given to the actual value in period t is the **smoothing constant** α, and the weight given to the forecast in period t is $1 - \alpha$. It turns out that the exponential smoothing forecast for any period is actually a weighted average of all the previous actual values of the time series. Let us illustrate by working with a time series involving only three periods of data: y_1, y_2, and y_3.

To initiate the calculations, we let $\hat{y}_1$ equal the actual value of the time series in period 1; that is, $\hat{y}_1 = y_1$. Hence, the forecast for period 2 is

$$\begin{aligned}\hat{y}_2 &= \alpha y_1 + (1-\alpha)\hat{y}_1 \\ &= \alpha y_1 + (1-\alpha) y_1 \\ &= y_1\end{aligned}$$

We see that the exponential smoothing forecast for period 2 is equal to the actual value of the time series in period 1.

The forecast for period 3 is

$$\hat{y}_3 = \alpha y_2 + (1 - \alpha)\hat{y}_2 = \alpha y_2 + (1 - \alpha) y_1$$

Finally, substituting this expression for $\hat{y}_3$ into the expression for $\hat{y}_4$, we obtain

$$\begin{aligned}\hat{y}_4 &= \alpha y_3 + (1 - \alpha)\hat{y}_3 \\ &= \alpha y_3 + (1 - \alpha)\left(\alpha y_2 + (1 - \alpha) y_1\right) \\ &= \alpha y_3 + \alpha(1 - \alpha) y_2 + (1 - \alpha)^2 y_1\end{aligned}$$

We now see that $\hat{y}_4$ is a weighted average of the first three time series values. The sum of the coefficients, or weights, for y_1, y_2, and y_3 equals 1. A similar argument can be made to show that, in general, any forecast $\hat{y}_{t+1}$ is a weighted average of all the t previous time series values.

Despite the fact that exponential smoothing provides a forecast that is a weighted average of all past observations, all past data do not need to be retained to compute the forecast for the next period. In fact, equation (9.7) shows that once the value for the smoothing constant α is selected, only two pieces of information are needed to compute the forecast

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

for period $t + 1$: y_t, the actual value of the time series in period t; and $\hat{y}_t$, the forecast for period t.

To illustrate the exponential smoothing approach to forecasting, let us again consider the gasoline sales time series in Table 9.1. As indicated previously, to initialize the calculations we set the exponential smoothing forecast for period 2 equal to the actual value of the time series in period 1. Thus, with $y_1 = 17$, we set $\hat{y}_2 = 17$ to initiate the computations. Referring to the time series data in Table 9.1, we find an actual time series value in period 2 of $y_2 = 21$. Thus, in period 2 we have a forecast error of $e_2 = 21 - 17 = 4$.

Continuing with the exponential smoothing computations using a smoothing constant of $\alpha = 0.2$, we obtain the following forecast for period 3:

$$\hat{y}_3 = 0.2y_2 + 0.8\hat{y}_2 = 0.2(21) + 0.8(17) = 17.8$$

Once the actual time series value in period 3, $y_3 = 19$, is known, we can generate a forecast for period 4 as follows:

$$\hat{y}_4 = 0.2y_3 + 0.8\hat{y}_3 = 0.2(19) + 0.8(17.8) = 18.04$$

Continuing the exponential smoothing calculations, we obtain the weekly forecast values shown in Table 9.10. Note that we have not shown an exponential smoothing forecast or a forecast error for week 1 because no forecast was made (we used actual sales for week 1 as the forecasted sales for week 2 to initialize the exponential smoothing process). For week 12, we have $y_{12} = 22$ and $\hat{y}_{12} = 18.48$. We can we use this information to generate a forecast for week 13:

$$\hat{y}_{13} = 0.2y_{12} + 0.8\hat{y}_{12} = 0.2(22) + 0.8(18.48) = 19.18$$

Thus, the exponential smoothing forecast of the amount sold in week 13 is 19.18, or 19,180 gallons of gasoline. With this forecast, the firm can make plans and decisions accordingly.

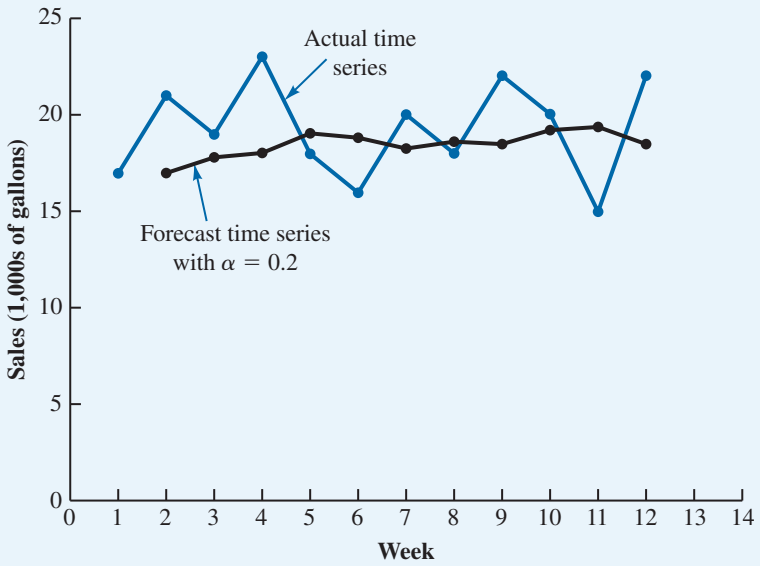
Figure 9.11 shows the time series plot of the actual and forecasted time series values. Note in particular how the forecasts smooth out the irregular or random fluctuations in the time series.

Table 9.10 Summary of the Exponential Smoothing Forecasts and Forecast Errors for the Gasoline Sales Time Series with Smoothing Constant $\alpha = 0.2$

Week	Time Series Value	Forecast	Forecast Error	Squared Forecast Error
1	17			
2	21	17.00	4.00	16.00
3	19	17.80	1.20	1.44
4	23	18.04	4.96	24.60
5	18	19.03	−1.03	1.06
6	16	18.83	−2.83	8.01
7	20	18.26	1.74	3.03
8	18	18.61	−0.61	0.37
9	22	18.49	3.51	12.32
10	20	19.19	0.81	0.66
11	15	19.35	−4.35	18.92
12	22	18.48	3.52	12.39
		Totals	10.92	98.80

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.11 Actual and Forecast Gasoline Time Series with Smoothing Constant $\alpha = 0.2$



DATA*file* gasoline

To show how Excel can be used for exponential smoothing, we again develop a forecast for the gasoline sales time series in Table 9.1. We use the file *gasoline,* which has the week in rows 2 through 13 of column A and the sales data for the 12 weeks in rows 2 through 13 of column B. We use $\alpha = 0.2$. The following steps can be used to produce a forecast.

Step 1. Click the **Data** tab in the Ribbon
Step 2. Click **Data Analysis** in the **Analyze** group
Step 3. When the **Data Analysis** dialog box appears (Figure 9.12), select **Exponential Smoothing** and click **OK**
Step 4. When the **Exponential Smoothing** dialog box appears (Figure 9.13):
Enter *B2:B13* in the **Input Range:** box
Enter *0.8* in the **Damping factor:** box
Enter *C2* in the **Output Range:** box
Click **OK**

Once you have completed this step, the exponential smoothing forecasts will appear in column C of the worksheet as shown in Figure 9.14. Note that the value we entered in the **Damping factor:** box is $1 - \alpha$; forecasts for other smoothing constants can be computed easily by entering a different value for $1 - \alpha$ in the **Damping factor:** box.

In the preceding exponential smoothing calculations, we used a smoothing constant of $\alpha = 0.2$. Although any value of α between 0 and 1 is acceptable, some values will yield more accurate forecasts than others. Insight into choosing a good value for α can be obtained by rewriting the basic exponential smoothing model as follows:

$$\begin{aligned}\hat{y}_{t+1} &= \alpha y_t + (1-\alpha)\hat{y}_t \\ &= \alpha y_t + \hat{y}_t - \alpha\hat{y}_t \\ &= \hat{y}_t + \alpha(y_t - \hat{y}_t) = \hat{y}_t + \alpha e_t\end{aligned}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.12 Data Analysis Dialog Box

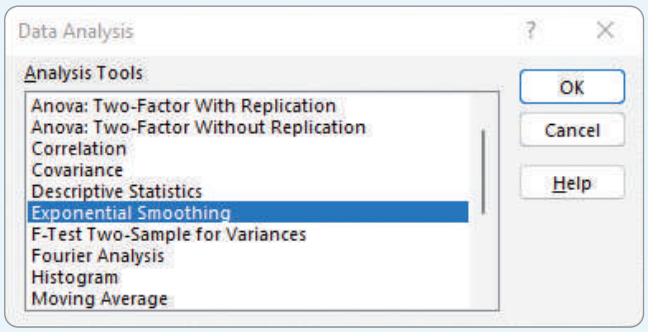


Figure 9.13 Exponential Smoothing Dialog Box

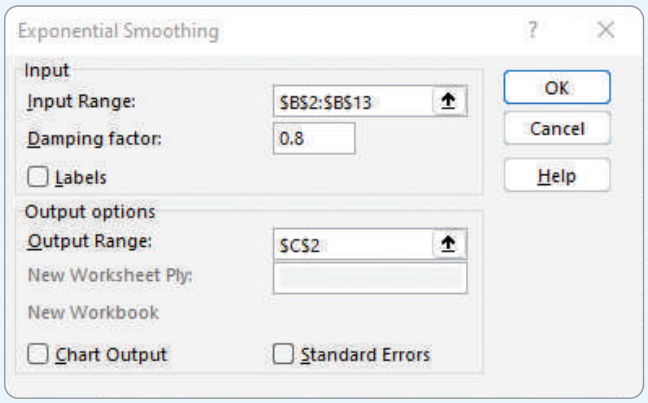


Figure 9.14 Excel Output for Exponential Smoothing Forecast for Gasoline Data

	A	B	C
1	**Week**	**Sales (1000s of gallons)**	
2	1	17	#N/A
3	2	21	17
4	3	19	17.8
5	4	23	18.04
6	5	18	19.032
7	6	16	18.8256
8	7	20	18.26048
9	8	18	18.60838
10	9	22	18.48671
11	10	20	19.18937
12	11	15	19.35149
13	12	22	18.48119

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Thus, the new forecast $\hat{y}_{t+1}$ is equal to the previous forecast $\hat{y}_t$ plus an adjustment, which is the smoothing constant α times the most recent forecast error, $e_t = y_t - \hat{y}_t$. In other words, the forecast in period $t + 1$ is obtained by adjusting the forecast in period t by a fraction of the forecast error from period t. If the time series contains substantial random variability, a small value of the smoothing constant is preferred. The reason for this choice is that if much of the forecast error is due to random variability, we do not want to overreact and adjust the forecasts too quickly. For a time series with relatively little random variability, a forecast error is more likely to represent a real change in the level of the series. Thus, larger values of the smoothing constant provide the advantage of quickly adjusting the forecasts to changes in the time series, thereby allowing the forecasts to react more quickly to changing conditions.

The criterion we will use to determine a desirable value for the smoothing constant α is the same as that proposed for determining the order or number of periods of data to include in the moving averages calculation; that is, we choose the value of α that minimizes the MSE. A summary of the MSE calculations for the exponential smoothing forecast of gasoline sales with $\alpha = 0.2$ is shown in Table 9.10. Note that there is one less squared error term than the number of time periods; this is because we had no past values with which to make a forecast for period 1. The value of the sum of squared forecast errors is 98.80; hence, MSE $= 98.80/11 = 8.98$. Would a different value of α provide better results in terms of a lower MSE value? Trial and error is often used to determine whether a different smoothing constant α can provide more accurate forecasts.

Nonlinear optimization can be used to identify the value of α that minimizes the MSE. Nonlinear optimization is discussed in Chapter 16.

Notes + Comments

1. Spreadsheet packages are effective tools for implementing exponential smoothing. With the time series data and the forecasting formulas in a spreadsheet you can use the MAE, MSE, and MAPE to evaluate different values of the smoothing constant α.
2. Moving averages and exponential smoothing provide the foundation for much of time series analysis, and many more sophisticated refinements of these methods have been developed. These include, but are not limited to, weighted moving averages, double moving averages, Brown's method for double exponential smoothing, and Holt-Winters exponential smoothing. Appendix 9.1 explains how to implement the Holt-Winters method using Excel Forecast Sheet.

9.4 Using Linear Regression Analysis for Forecasting

In Chapter 8, we discuss linear regression models in more detail.

Regression analysis is a statistical technique that can be used to develop a mathematical equation showing how variables are related. In regression terminology, the variable that is being predicted is called the *dependent* (or *response*) *variable*, and the variable or variables being used to predict the value of the dependent variable are called the *independent* (or *predictor*) *variables*. In this section, we will show how to use linear regression analysis to develop forecasts for a time series that has a trend, a seasonal pattern, and both a trend and a seasonal pattern. We will also show how to use linear regression analysis to develop forecast models that include causal variables.

Linear Trend Projection

We now consider forecasting methods that are appropriate for time series that exhibit trend patterns and show how linear regression analysis can be used to forecast a time series with a linear trend. In Section 9.1, we used the bicycle sales time series in Table 9.3 to illustrate a time series with a trend pattern. Let us now use this time series to illustrate how linear regression analysis can be used to forecast a time series with a linear trend. Although the time series plot in Figure 9.3 shows some up-and-down movement over the past 10 years,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

we might agree that a linear trendline provides a reasonable approximation of the long-run movement in the series. We can use linear regression analysis to develop such a linear trendline for the bicycle sales time series.

Because simple linear regression analysis yields the linear relationship between the independent variable and the dependent variable that minimizes the MSE, we can use this approach to find a best-fitting line to a set of data that exhibits a linear trend. In finding a linear trend, the variable to be forecasted (y_t, the actual value of the time series in period t) is the dependent variable and the trend variable (time period t) is the independent variable. We will use the following notation for our linear trendline:

$$\hat{y}_t = b_0 + b_1 t \tag{9.8}$$

where

$\hat{y}_t$ = forecast of sales in period t
t = time period
b_0 = the y-intercept of the linear trendline
b_1 = the slope of the linear trendline

In equation (9.8), the time variable begins at $t = 1$, corresponding to the first time series observation (year 1 for the bicycle sales time series). The time variable then continues until $t = n$, corresponding to the most recent time series observation (year 10 for the bicycle sales time series). Thus, the bicycle sales time series $t = 1$ corresponds to the oldest time series value, and $t = 10$ corresponds to the most recent year.

Excel can be used to compute the estimated intercept b_0 and slope b_1. The Excel output for a regression analysis of the bicycle sales data is provided in Figure 9.15.

We see in this output that the estimated intercept b_0 is 20.4 (shown in cell B17) and the estimated slope b_1 is 1.1 (shown in cell B18). Thus,

$$\hat{y}_t = 20.4 + 1.1t \tag{9.9}$$

is the linear regression equation for the linear trend component for the bicycle sales time series. The slope of 1.1 in this trend equation indicates that over the past 10 years the firm has experienced an average growth in sales of about 1,100 units per year. If we assume that the past 10-year trend in sales is a good indicator for the future, we can use equation (9.9) to project the trend component of the time series. For example, substituting $t = 11$ into equation (9.9) yields next year's trend projection, $\hat{y}_{11}$:

$$\hat{y}_{11} = 20.4 + 1.1(11) = 32.5$$

Thus, the linear trend model yields a sales forecast of 32,500 bicycles for the next year.

We can also use the trendline to forecast sales farther into the future. Using equation (9.9), we develop annual forecasts of bicycle sales for two and three years into the future as follows:

$$\hat{y}_{12} = 20.4 + 1.1(12) = 33.6$$
$$\hat{y}_{13} = 20.4 + 1.1(13) = 34.7$$

The forecasted value increases by 1,100 bicycles in each year.

Note that to produce a forecast of the value of the dependent variable y for period t in this example, we are not using values of the dependent variable from previous periods (e.g., y_{t-1}, y_{t-2}, y_{t-3},…) as independent variables. Thus, $k = 0$ in equations (9.3)–(9.5) to calculate the MAE, MSE, and MAPE.

We can also use more complex linear regression models to fit nonlinear trends. For example, to generate a forecast of a time series with a curvilinear trend, we could include t^2 and t^3 as independent variables in our model, and the estimated linear regression equation would become

$$\hat{y}_t = b_0 + b_1 t + b_2 t^2 + b_3 t^3$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.15 Excel Simple Linear Regression Output for Trendline Model for Bicycle Sales Data

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.874526167							
5	R Square	0.764796016							
6	Adjusted R Square	0.735395518							
7	Standard Error	1.958953802							
8	Observations	10							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	99.825	99.825	26.01302932	0.000929509			
13	Residual	8	30.7	3.8375					
14	Total	9	130.525						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	20.4	1.338220211	15.24412786	3.39989E-07	17.31405866	23.48594134	15.90975286	24.89024714
18	Year	1.1	0.215673715	5.100296983	0.000929509	0.60265552	1.59734448	0.376331148	1.823668852

Because autoregressive models typically violate the conditions necessary for inference in least squares regression, you must be careful when testing hypotheses or estimating confidence intervals in autoregressive models. There are special methods for constructing autoregressive models, but they are beyond the scope of this book.

Another type of linear regression-based forecasting model occurs whenever all the independent variables are previous values of the same time series. For example, if the time series values are denoted by $y_1, y_2, \ldots, y_n$, we might try to find an estimated linear regression equation relating y_t to the most recent time series values, y_{t-1}, y_{t-2}, and so on. If we use the actual values of the time series for the three most recent periods as independent variables, the estimated regression equation would be

$$\hat{y}_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + b_3 y_{t-3}$$

Linear regression models such as this in which the independent variables are previous values of the time series are referred to as **autoregressive models**.

Seasonality Without Trend

To the extent that seasonality exists, we need to incorporate it into our forecasting models to ensure accurate forecasts. We begin by considering a seasonal time series with no trend and then, in the next section, we discuss how to model seasonality with a linear trend. Let us consider again the data from Table 9.5, the number of umbrellas sold at a clothing store over the past five years. As we see in the time series plot provided in Figure 9.5, the data do not suggest any long-term trend in sales. In fact, unless you look carefully at the data, you might conclude that the data follow a horizontal pattern with random fluctuation and that single exponential smoothing could be used to forecast sales. However, closer inspection of the time series plot reveals a pattern in the fluctuations. The first and third quarters have moderate sales, the second quarter the highest sales, and the fourth quarter the lowest sales. Thus, we conclude that a quarterly seasonal pattern is present.

Categorical variables are covered in more detail in Chapter 8.

We can model a time series with a seasonal pattern by treating the season as a dummy variable. Categorical variables are variables used to categorize observations of data, and $k - 1$ dummy variables are required to model a categorical variable that has k levels. Thus,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

we need three dummy variables to model four seasons. For instance, in the umbrella sales time series, the quarter to which each observation corresponds is treated as a season; it is a categorical variable with four levels: quarter 1, quarter 2, quarter 3, and quarter 4. Thus, to model the seasonal effects in the umbrella time series we need $4 - 1 = 3$ dummy variables. The three dummy variables can be coded as follows:

$$\text{Qtr1}_t = \begin{cases} 1 \text{ if period } t \text{ is quarter 1} \\ 0 \text{ otherwise} \end{cases}$$

$$\text{Qtr2}_t = \begin{cases} 1 \text{ if period } t \text{ is quarter 2} \\ 0 \text{ otherwise} \end{cases}$$

$$\text{Qtr3}_t = \begin{cases} 1 \text{ if period } t \text{ is quarter 3} \\ 0 \text{ otherwise} \end{cases}$$

Using $\hat{y}_t$ to denote the forecasted value of sales for period t, the general form of the equation relating the number of umbrellas sold to the quarter the sales take place is as follows:

$$\hat{y}_t = b_0 + b_1\text{Qtr1}_t + b_2\text{Qtr2}_t + b_3\text{Qtr3}_t \quad \textbf{(9.10)}$$

Note that the fourth quarter will be denoted by setting all three dummy variables to 0.

Table 9.11 shows the umbrella sales time series with the coded values of the dummy variables shown. We can use a multiple linear regression model to find the values of b_0, b_1, b_2, and b_3 that minimize the sum of squared errors. For this multiple linear regression model, y_t is the dependent variable, and the quarterly dummy variables Qtr1_t, Qtr2_t, and Qtr3_t are the independent variables.

Using the data in Table 9.11 and linear regression analysis, we obtain the following equation:

$$\hat{y}_t = 95.0 + 29.0\text{Qtr1}_t + 57.0\text{Qtr2}_t + 26.0\text{Qtr3}_t \quad \textbf{(9.11)}$$

We can use equation (9.11) to forecast sales of every quarter for the next year:

$$\begin{aligned} \text{Quarter1}: \text{Sales} &= 95.0 + 29.0(1) + 57.0(0) + 26.0(0) = 124 \\ \text{Quarter2}: \text{Sales} &= 95.0 + 29.0(0) + 57.0(1) + 26.0(0) = 152 \\ \text{Quarter3}: \text{Sales} &= 95.0 + 29.0(0) + 57.0(0) + 26.0(1) = 121 \\ \text{Quarter4}: \text{Sales} &= 95.0 + 29.0(0) + 57.0(0) + 26.0(0) = 95 \end{aligned}$$

It is interesting to note that we could have obtained the quarterly forecasts for the next year by simply computing the average number of umbrellas sold in each quarter. Nonetheless, for more complex problem situations, such as dealing with a time series that has both trend and seasonal effects, this simple averaging approach will not work.

Seasonality with Trend

We now consider situations for which the time series contains both seasonal effects and a linear trend by showing how to forecast the quarterly sales of smartphones introduced in Section 9.1. The data for the smartphone time series are shown in Table 9.6. The time series plot in Figure 9.6 indicates that sales are lowest in the second quarter of each year and increase in quarters 3 and 4. Thus, we conclude that a seasonal pattern exists for smartphone sales. However, the time series also has an upward linear trend that will need to be accounted for in order to develop accurate forecasts of quarterly sales. This is easily done by combining the dummy variable approach for handling seasonality with the approach for handling a linear trend discussed earlier in this section.

The general form of the linear regression equation for modeling both the quarterly seasonal effects and the linear trend in the smartphone time series is

$$\hat{y}_t = b_0 + b_1\text{Qtr1}_t + b_2\text{Qtr2}_t + b_3\text{Qtr3}_t + b_4 t \quad \textbf{(9.12)}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 9.11 Umbrella Sales Time Series with Dummy Variables

Period	Year	Quarter	Qtr1	Qtr2	Qtr3	Sales
1	1	1	1	0	0	125
2		2	0	1	0	153
3		3	0	0	1	106
4		4	0	0	0	88
5	2	1	1	0	0	118
6		2	0	1	0	161
7		3	0	0	1	133
8		4	0	0	0	102
9	3	1	1	0	0	138
10		2	0	1	0	144
11		3	0	0	1	113
12		4	0	0	0	80
13	4	1	1	0	0	109
14		2	0	1	0	137
15		3	0	0	1	125
16		4	0	0	0	109
17	5	1	1	0	0	130
18		2	0	1	0	165
19		3	0	0	1	128
20		4	0	0	0	96

where

$\hat{y}_t$ = forecast of sales in period t
Qtr1_t = 1 if time period t corresponds to the first quarter of the year; 0 otherwise
Qtr2_t = 1 if time period t corresponds to the second quarter of the year; 0 otherwise
Qtr3_t = 1 if time period t corresponds to the third quarter of the year; 0 otherwise
t = time period (quarter)

For this linear regression model, y_t is the dependent variable and the quarterly dummy variables Qtr1_t, Qtr2_t, and Qtr3_t and the time period t are the independent variables.

Table 9.12 shows the revised smartphone sales time series that includes the coded values of the dummy variables and the time period t. Using the data in Table 9.12 with the linear regression model that includes both the seasonal and trend components, we obtain the following equation that minimizes our sum of squared errors:

$$\hat{y}_t = 6.07 - 1.36\text{Qtr1}_t - 2.03\text{Qtr2}_t - 0.304\text{Qtr3}_t + 0.146t \quad \textbf{(9.13)}$$

We can now use equation (9.13) to forecast quarterly sales for the next year. Next year is year 5 for the smartphone sales time series, that is, time periods 17, 18, 19, and 20.

Forecast for time period 17 (quarter 1 in year 5):

$$\hat{y}_{17} = 6.07 - 1.36(1) - 2.03(0) - 0.304(0) + 0.146(17) = 7.19$$

Forecast for time period 18 (quarter 2 in year 5):

$$\hat{y}_{18} = 6.07 - 1.36(0) - 2.03(1) - 0.304(0) + 0.146(18) = 6.67$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 9.12 Smartphone Sales Time Series with Dummy Variables and Time Period

Period	Year	Quarter	Qtr1	Qtr2	Qtr3	Sales (1,000s)
1	1	1	1	0	0	4.8
2		2	0	1	0	4.1
3		3	0	0	1	6.0
4		4	0	0	0	6.5
5	2	1	1	0	0	5.8
6		2	0	1	0	5.2
7		3	0	0	1	6.8
8		4	0	0	0	7.4
9	3	1	1	0	0	6.0
10		2	0	1	0	5.6
11		3	0	0	1	7.5
12		4	0	0	0	7.8
13	4	1	1	0	0	6.3
14		2	0	1	0	5.9
15		3	0	0	1	8.0
16		4	0	0	0	8.4

Forecast for time period 19 (quarter 3 in year 5):

$$\hat{y}_{19} = 6.07 - 1.36(0) - 2.03(0) - 0.304(1) + 0.146(19) = 8.54$$

Forecast for time period 20 (quarter 4 in year 5):

$$\hat{y}_{20} = 6.07 - 1.36(0) - 2.03(0) - 0.304(0) + 0.146(20) = 8.99$$

Thus, accounting for the seasonal effects and the linear trend in smartphone sales, the estimates of quarterly sales in year 5 are 7,190; 6,670; 8,540; and 8,990.

The dummy variables in the equation actually provide four equations, one for each quarter. For instance, if time period t corresponds to quarter 1, the estimate of quarterly sales is

$$\text{Quarter 1: Sales} = 6.07 - 1.36(1) - 2.03(0) - 0.304(0) + 0.146t = 4.71 + 0.146t$$

Similarly, if time period t corresponds to quarters 2, 3, and 4, the estimates of quarterly sales are as follows:

$$\text{Quarter 2: Sales} = 6.07 - 1.36(0) - 2.03(1) - 0.304(0) + 0.146t = 4.04 + 0.146t$$
$$\text{Quarter 3: Sales} = 6.07 - 1.36(0) - 2.03(0) - 0.304(1) + 0.146t = 5.77 + 0.146t$$
$$\text{Quarter 4: Sales} = 6.07 - 1.36(0) - 2.03(0) - 0.304(0) + 0.146t = 6.07 + 0.146t$$

The slope of the trendline for each quarterly forecast equation is 0.146, indicating a consistent growth in sales of about 146 phones per quarter. The only difference in the four equations is that they have different intercepts.

In the smartphone sales example, we showed how dummy variables can be used to account for the quarterly seasonal effects in the time series. Because there were four levels of seasonality, three dummy variables were required. However, many businesses use monthly rather than quarterly forecasts. For monthly data, season is a categorical variable

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

with 12 levels, and thus $12 - 1 = 11$ dummy variables are required to capture monthly seasonal effects. For example, the 11 dummy variables could be coded as follows:

$$\text{Month1}_t = \begin{cases} 1 \text{ if period } t \text{ is January} \\ 0 \text{ otherwise} \end{cases}$$

$$\text{Month2}_t = \begin{cases} 1 \text{ if period } t \text{ is February} \\ 0 \text{ otherwise} \end{cases}$$

$$\vdots$$

$$\text{Month11}_t = \begin{cases} 1 \text{ if period } t \text{ is November} \\ 0 \text{ otherwise} \end{cases}$$

Other than this change, the approach for handling seasonality remains the same. Time series data collected at other intervals can be handled in a similar manner.

Using Linear Regression Analysis as a Causal Forecasting Method

The methods discussed for estimating linear trends and seasonal effects make use of patterns in historical values of the variable to be forecast; these methods are classified as time series methods because they rely on past values of the variable to be forecast when developing the model. However, the relationship of the variable to be forecast with other variables may also be used to develop a forecasting model. Generally such models include only variables that are believed to cause changes in the variable to be forecast, such as the following:

- Advertising expenditures when sales are to be forecast.
- The mortgage rate when new housing construction is to be forecast.
- Grade point average when starting salaries for recent college graduates are to be forecast.
- The price of a product when the demand for the product is to be forecast.
- The value of the Dow Jones Industrial Average when the value of an individual stock is to be forecast.
- Daily high temperature when electricity usage is to be forecast.

Because these variables are used as independent variables when we believe they cause changes in the value of the dependent variable, forecasting models that include such variables as independent variables are referred to as **causal models**. It is important to note here that the forecasting model provides evidence only of association between an independent variable and the variable to be forecast. The model does not provide evidence of a causal relationship between an independent variable and the variable to be forecast; the conclusion that a causal relationship exists must be based on practical experience.

To illustrate how linear regression analysis is used as a causal forecasting method, we consider the sales forecasting problem faced by Armand's Pizza Parlors, a chain of Italian restaurants doing business in a five-state area. Historically, the most successful locations have been near college campuses. The managers believe that quarterly sales for these restaurants (denoted by y) are related positively to the size of the student population (denoted by x); that is, restaurants near campuses with a large population tend to generate more sales than those located near campuses with a small population.

Using linear regression analysis, we can develop an equation showing how the dependent variable y is related to the independent variable x. This equation can then be used to forecast quarterly sales for restaurants located near college campuses given the size of the student population. This is particularly helpful for forecasting sales for new restaurant locations. For instance, suppose that management wants to forecast sales for a new restaurant that it is considering opening near a college campus. Because no historical data are available on sales for a new restaurant, Armand's cannot use time series data to develop the forecast. However, as we will now illustrate, regression analysis can still be used to forecast quarterly sales for this new location.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To develop the equation relating quarterly sales to the size of the student population, Armand's collected data from a sample of 10 of its restaurants located near college campuses. These data are summarized in Table 9.13. For example, restaurant 1, with $y = 58$ and $x = 2$, had \$58,000 in quarterly sales and is located near a campus with 2,000 students. Figure 9.16 shows a scatter chart of the data presented in Table 9.13, with the size of the student population shown on the horizontal axis and quarterly sales shown on the vertical axis.

What preliminary conclusions can we draw from Figure 9.16? Sales appear to be higher at locations near campuses with larger student populations. Also, it appears that the relationship between the two variables can be approximated by a straight line. In Figure 9.17, we can draw a straight line through the data that appears to provide a good linear approximation of the relationship between the variables. Observe that the relationship is not perfect. Indeed, few, if any, of the data fall exactly on the line. However, if we can develop the mathematical expression for this line, we may be able to use it to forecast the value of y corresponding to each possible value of x. The resulting equation of the line is called the estimated linear regression equation.

Using the least squares method of estimation, the estimated linear regression equation is

$$\hat{y}_i = b_0 + b_1 x_i \tag{9.14}$$

where

$\hat{y}_i$ = estimated value of the dependent variable (quarterly sales) for the ith observation
b_0 = intercept of the estimated regression equation
b_1 = slope of the estimated regression equation
x_i = value of the independent variable (student population) for the ith observation

The Excel output for a simple linear regression analysis of the Armand's Pizza data is provided in Figure 9.18.

We see in this output that the estimated intercept b_0 is 60 and the estimated slope b_1 is 5. Thus, the estimated linear regression equation is

$$\hat{y}_i = 60 + 5x_i$$

The slope of the estimated linear regression equation ($b_1 = 5$) is positive, implying that, as student population increases, quarterly sales increase. In fact, we can conclude (because sales are measured in thousands of dollars and student population in thousands) that an increase in the student population of 1,000 is associated with an increase of \$5,000 in

Table 9.13 Student Population and Quarterly Sales Data for 10 Armand's Pizza Parlors

Restaurant	Student Population (1,000s)	Quarterly Sales (\$1,000s)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

DATA*file*
armands

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.16 Scatter Chart of Student Population and Quarterly Sales for Armand's Pizza Parlors

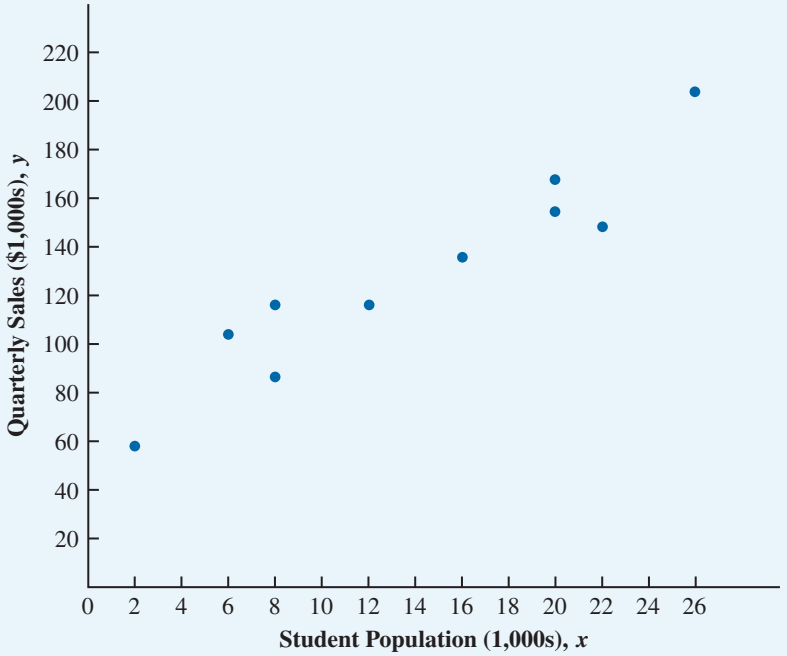
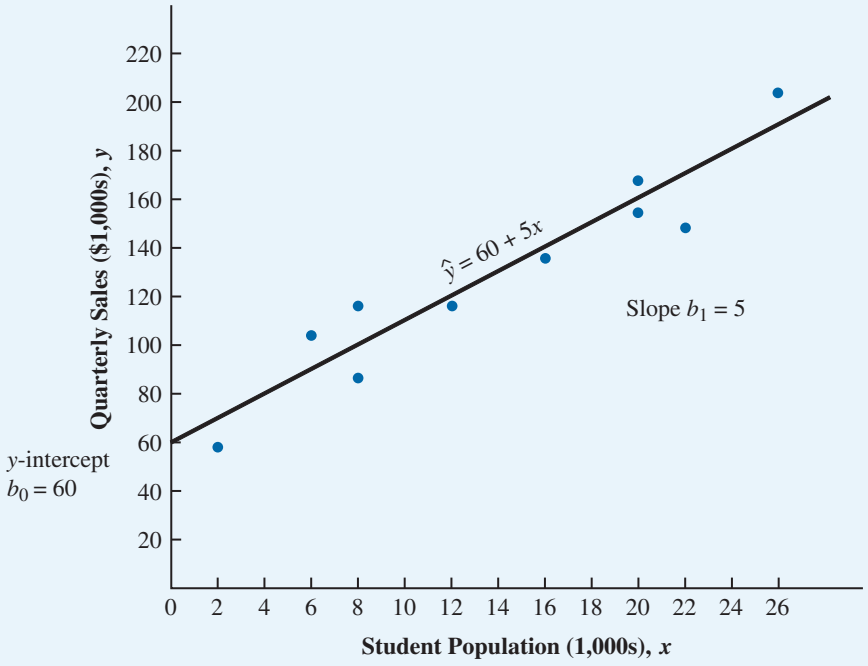


Figure 9.17 Graph of the Estimated Linear Regression Equation for Armand's Pizza Parlors: $y = 60 + 5x$



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.18 Excel Simple Linear Regression Output for Armand's Pizza Parlors

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.950122955							
5	R Square	0.90273363							
6	Adjusted R Square	0.890575334							
7	Standard Error	13.82931669							
8	Observations	10							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	14200	14200	74.24836601	2.54887E-05			
13	Residual	8	1530	191.25					
14	Total	9	15730						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	60	9.22603481	6.503335532	0.000187444	38.72472558	81.27527442	29.04307968	90.95692032
18	Student Population (1000s)	5	0.580265238	8.616749156	2.54887E-05	3.661905962	6.338094038	3.052985371	6.947014629

Note that the values of the independent variable range from 2,000 to 26,000; thus, as discussed in Chapter 8, the y-intercept in such cases is an extrapolation of the linear regression line and must be interpreted with caution.

expected quarterly sales; that is, quarterly sales are expected to increase by $5 per student. The estimated y-intercept b_0 tells us that if the student population for the location of an Armand's pizza parlor was 0 students, we would expect sales of $60,000.

If we believe that the least squares estimated linear regression equation adequately describes the relationship between x and y, using the estimated linear regression equation to forecast the value of y for a given value of x seems reasonable. For example, if we wanted to forecast quarterly sales for a new restaurant to be located near a campus with 16,000 students, we would compute as follows:

$$\hat{y} = 60 + 5(16)$$
$$= 140$$

Hence, we would forecast quarterly sales of $140,000.

Combining Causal Variables with Trend and Seasonality Effects

The value of an independent variable from the prior period is referred to as a lagged variable.

Linear regression models are very flexible and can incorporate both causal variables and time series effects. Suppose we had a time series of several years of quarterly sales data and advertising expenditures for a single Armand's restaurant. If we suspected that sales were related to advertising expenditures and that sales showed trend and seasonal effects, we could incorporate each into a single model by combining the approaches we have outlined. If we believe that the effect of advertising is not immediate, we might also try to find a relationship between sales in period t and advertising in the previous period, $t - 1$.

Multiple linear regression analysis also can be applied in these situations if additional data for other independent variables are available. For example, suppose that the management of Armand's Pizza Parlors also believes that the number of competitors near the college campus is related to quarterly sales. Intuitively, management believes that restaurants located near campuses with fewer competitors generate more sales revenue than those located near campuses with more competitors. With additional data, multiple linear

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

regression analysis could be used to develop an equation relating quarterly sales to the size of the student population and the number of competitors.

Considerations in Using Linear Regression in Forecasting

Although linear regression analysis allows for the estimation of complex forecasting models, we must be cautious about using such models and guard against the potential for overfitting our model to the sample data. Spyros Makridakis, a noted forecasting expert, conducted research showing that simple techniques usually outperform more complex procedures for short-term forecasting. Using a more sophisticated procedure will not guarantee better forecasts. However, many research studies, including those done by Makridakis, have also shown that quantitative forecasting models such as those presented in this chapter commonly outperform qualitative forecasts made by "experts." Thus, there is good reason to use quantitative forecasting methods whenever data are available.

Whether a linear regression approach provides a good forecast depends largely on how well we are able to identify and obtain data for independent variables that are closely related to the time series. Generally, during the development of an estimated linear regression equation, we will want to consider many possible sets of independent variables. Thus, part of the linear regression analysis procedure should focus on the selection of the set of independent variables that provides the best forecasting model.

Notes + Comments

Many different software packages can be used to estimate linear regression models. Section 8.4 in this textbook explains how Excel's Regression tool can be used to perform linear regression analysis. The online appendixes available with this text demonstrate the use of several popular statistical packages to estimate linear regression models.

9.5 Determining the Best Forecasting Model to Use

Given the variety of forecasting models and approaches, the obvious question is, "For a given forecasting study, how does one choose an appropriate model?" As discussed throughout this text, it is always a good idea to get descriptive statistics on the data and graph the data so that they can be visually inspected. In the case of time series data, a visual inspection can indicate whether seasonality appears to be a factor and whether a linear or nonlinear trend seems to exist. For causal modeling, scatter charts can indicate whether strong linear or nonlinear relationships exist between the independent and dependent variables. If certain relationships appear totally random, this may lead you to exclude these variables from the model.

As in linear regression analysis, you may be working with large data sets when generating a forecasting model. In such cases, it is recommended to divide your data into training and validation sets. For example, you might have five years of monthly data available to produce a time series forecast. You could use the first three years of data as a training set to estimate a model or a collection of models that appear to provide good forecasts. You might develop exponential smoothing models and linear regression models for the training set. You could then use the last two years as a validation set to assess and compare the models' performances. Based on the errors produced by the different models for the validation set, you could ultimately pick the model that minimizes some forecast error measure, such as MAE, MSE, or MAPE. However, you must exercise caution in using the older portion of a time series for the training set and the more recent portion of the time series as the validation set. If the behavior of the time series has changed recently, the older portion of the time series may no longer show patterns similar to the more recent values of the time series, and a forecasting model based on such data will not perform well.

Some software packages try many different forecasting models on time series data (those included in this chapter and more) and report back optimal model parameters and error measures for each model tested. Although some of these software packages will even automatically select the best model to use, ultimately the user should decide which model to use going

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

forward based on a combination of the software output and the user's knowledge about the domain where the forecast model will be applied.

Summary

This chapter provided an introduction to the basic methods of time series analysis and forecasting. First, we showed that to explain the behavior of a time series, it is often helpful to graph the time series and identify whether trend, seasonal, and/or cyclical components are present in the time series. The methods we have discussed are based on assumptions about which of these components are present in the time series.

We discussed how smoothing methods can be used to forecast a time series that exhibits no significant trend, seasonal, or cyclical effect. The moving average approach consists of computing an average of past data values and then using that average as the forecast for the next period. In the exponential smoothing method, a weighted average of past time series values is used to compute a forecast.

For time series that have only a long-term trend, we showed how linear regression analysis could be used to make trend projections. For time series with seasonal influences, we showed how to incorporate the seasonality for more accurate forecasts. We described how linear regression analysis can be used to develop causal forecasting models that relate values of the variable to be forecast (the dependent variable) to other independent variables that are believed to explain (cause) the behavior of the dependent variable. Finally, we have provided guidance on how to select an appropriate model from the models discussed in this chapter.

Glossary

Autoregressive model A regression model in which a regression relationship based on past time series values is used to predict the future time series values.
Causal models Forecasting methods that relate a time series to other variables that are believed to explain or cause its behavior.
Cyclical pattern The component of the time series that results in periodic above-trend and below-trend behavior of the time series lasting more than one year.
Exponential smoothing A forecasting technique that uses a weighted average of past time series values as the forecast.
Forecast error The amount by which the forecasted value $\hat{y}_t$ differs from the observed value y_t, denoted by $e_t = y_t - \hat{y}_t$.
Forecasts A prediction of future values of a time series.
Mean absolute error (MAE) A measure of forecasting accuracy; the average of the absolute values of the forecast errors. Also referred to as mean absolute deviation (MAD).
Mean absolute percentage error (MAPE) A measure of the accuracy of a forecasting method; the average of the absolute values of the errors as a percentage of the corresponding forecast values.
Mean squared error (MSE) A measure of the accuracy of a forecasting method; the average of the sum of the squared differences between the forecast values and the actual time series values.
Moving average method A method of forecasting or smoothing a time series that uses the average of the most recent n data values in the time series as the forecast for the next period.
Naïve forecasting method A forecasting technique that uses the value of the time series from the most recent period as the forecast for the current period.
Seasonal pattern The component of the time series that shows a periodic pattern over one year or less.
Smoothing constant A parameter of the exponential smoothing model that provides the weight given to the most recent time series value in the calculation of the forecast value.
Stationary time series A time series whose statistical properties are independent of time.
Time series A set of observations on a variable measured at successive points in time or over successive periods of time.
Trend The long-run shift or movement in the time series observable over several periods of time.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Problems

1. **Measuring the Forecast Accuracy of the Naïve Method.** Consider the following time series data:

Week	1	2	3	4	5	6
Value	18	13	16	11	17	14

Using the naïve method (most recent value) as the forecast for the next week, compute each of the following. LO 1, 4
   a. Mean absolute error
   b. Mean squared error
   c. Mean absolute percentage error
   d. Forecast for week 7

2. **Measuring the Forecast Accuracy of the Average of All Historical Data.** Refer to the time series data in Problem 1. Using the average of all the historical data as a forecast for the next period, compute the following. LO 1, 4
   a. Mean absolute error
   b. Mean squared error
   c. Mean absolute percentage error
   d. Forecast for week 7

3. **Comparing the Forecast Accuracy of the Naïve Method and the Average of All Historical Data.** Problems 1 and 2 used different forecasting methods. Which method appears to provide the more accurate forecasts for the historical data? Explain. LO 2

4. **Measuring the Forecast Accuracy for Monthly Data.** Consider the following time series data. LO 1, 2, 4

Month	1	2	3	4	5	6	7
Value	24	13	20	12	19	23	15

   a. Compute MSE using the most recent value as the forecast for the next period. What is the forecast for month 8?
   b. Compute MSE using the average of all the data available as the forecast for the next period. What is the forecast for month 8?
   c. Which method appears to provide the better forecast?

5. **Forecasting Weekly Data.** Consider the following time series data. LO 1, 2, 3, 5

Week	1	2	3	4	5	6
Value	18	13	16	11	17	14

   a. Construct a time series plot. What type of pattern exists in the data?
   b. Develop a three-week moving average for this time series. Compute MSE and a forecast for week 7.
   c. Use $\alpha = 0.2$ to compute the exponential smoothing values for the time series. Compute MSE and a forecast for week 7.
   d. Compare the three-week moving average forecast with the exponential smoothing forecast using $\alpha = 0.2$. Which appears to provide the better forecast based on MSE? Explain.
   e. Use trial and error to find a value of the exponential smoothing coefficient α that results in a smaller MSE than what you calculated for $\alpha = 0.2$.

6. **Forecasting Weekly Data.** Consider the following time series data. LO 1, 2, 3, 5

Week	1	2	3	4	5	6	7
Value	24	13	20	12	19	23	15

   a. Construct a time series plot. What type of pattern exists in the data?
   b. Develop a three-week moving average for this time series. Compute MSE and a forecast for week 8.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Use $\alpha = 0.2$ to compute the exponential smoothing values for the time series. Compute MSE and a forecast for week 8.
d. Compare the three-week moving average forecast with the exponential smoothing forecast using $\alpha = 0.2$. Which appears to provide the better forecast based on MSE?
e. Use trial and error to find a value of the exponential smoothing coefficient α that results in a smaller MSE than what you calculated for $\alpha = 0.2$.

DATA*file* gasoline

7. **Forecasting Gasoline Sales with Moving Averages.** Refer to the gasoline sales time series data in Table 9.1. LO 1, 2, 5
a. Compute four-week and five-week moving averages for the time series.
b. Compute the MSE for the four-week and five-week moving average forecasts.
c. What appears to be the best number of weeks of past data (three, four, or five) to use in the moving average computation? Recall that the MSE for the three-week moving average is 10.22.

DATA*file* gasoline

8. **Forecasting Gasoline Sales with Exponential Smoothing.** With the gasoline time series data from Table 9.1, show the exponential smoothing forecasts using $\alpha = 0.1$. LO 1, 2
a. Applying the MSE measure of forecast accuracy, would you prefer a smoothing constant of $\alpha = 0.1$ or $\alpha = 0.2$ for the gasoline sales time series?
b. Are the results the same if you apply MAE as the measure of accuracy?
c. What are the results if MAPE is used?

9. **Comparing Gasoline Sales Forecasts with Moving Averages and Exponential Smoothing.** With a smoothing constant of $\alpha = 0.2$, equation (9.7) shows that the forecast for week 13 of the gasoline sales data from Table 9.1 is given by $\hat{y}_{13} = 0.2y_{12} + 0.8\hat{y}_{12}$. However, the forecast for week 12 is given by $\hat{y}_{12} = 0.2y_{11} + 0.8\hat{y}_{11}$. Thus, we could combine these two results to show that the forecast for week 13 can be written as

$$\hat{y}_{13} = 0.2y_{12} + 0.8(0.2y_{11} + 0.8\hat{y}_{11}) = 0.2y_{12} + 0.16y_{11} + 0.64\hat{y}_{11}$$

Use this equation to answer the following questions. LO 5
a. Making use of the fact that $\hat{y}_{11} = 0.2y_{10} + 0.8\hat{y}_{10}$ (and similarly for $\hat{y}_{10}$ and $\hat{y}_9$), continue to expand the expression for $\hat{y}_{13}$ until it is written in terms of the past data values y_{12}, y_{11}, y_{10}, y_9, y_8, and the forecast for period 8, $\hat{y}_8$.
b. Refer to the coefficients or weights for the past values y_{12}, y_{11}, y_{10}, y_9, and y_8. What observation can you make about how exponential smoothing weights past data values in arriving at new forecasts? Compare this weighting pattern with the weighting pattern of the moving averages method.

10. **Demand for Dairy Products.** United Dairies, Inc. supplies milk to several independent grocers throughout Dade County, Florida. Managers at United Dairies want to develop a forecast of the number of half gallons of milk sold per week. Sales data for the past 24 weeks are as follows. LO 1, 3, 5

DATA*file* uniteddairies

Week	Sales	Week	Sales
1	2,750	13	3,150
2	3,100	14	3,520
3	3,250	15	3,640
4	2,800	16	3,190
5	2,900	17	3,310
6	3,050	18	3,450
7	3,300	19	3,700
8	3,100	20	3,490
9	2,950	21	3,350
10	3,000	22	3,380
11	3,200	23	3,600
12	3,150	24	3,540

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Construct a time series plot. What type of pattern exists in the data?
b. Use exponential smoothing with $\alpha = 0.4$ to develop a forecast of demand for week 25. What is the resulting MSE?

DATA*file* hawkins

11. **On-Time Shipments.** For the Hawkins Company, the monthly percentages of all shipments received on time over the past 12 months are 80, 82, 84, 83, 83, 84, 85, 84, 82, 83, 84, and 83. **LO 1, 2, 3, 5**
a. Construct a time series plot. What type of pattern exists in the data?
b. Compare a three-month moving average forecast with an exponential smoothing forecast for $\alpha = 0.2$. Which provides the better forecasts using MSE as the measure of model accuracy?
c. What is the forecast for the next month?

DATA*file* tripleabond

12. **Bond Interest Rates.** Corporate triple-A bond interest rates for 12 consecutive months are as follows. **LO 1, 2, 3, 5**

9.5 9.3 9.4 9.6 9.8 9.7 9.8 10.5 9.9 9.7 9.6 9.6

a. Construct a time series plot. What type of pattern exists in the data?
b. Develop three-month and four-month moving averages for this time series. Does the three-month or the four-month moving average provide the better forecasts based on MSE? Explain.
c. What is the moving average forecast for the next month?

DATA*file* alabama

13. **Building Contracts.** The values of Alabama building contracts (in millions of dollars) for a 12-month period are as follows. **LO 1, 3, 5**

240 350 230 260 280 320 220 310 240 310 240 230

a. Construct a time series plot. What type of pattern exists in the data?
b. Compare a three-month moving average forecast with an exponential smoothing forecast. Use $\alpha = 0.2$. Which provides the better forecasts based on MSE?
c. What is the forecast for the next month using exponential smoothing with $\alpha = 0.2$?

14. **Sales Forecasts.** The following time series shows the sales of a particular product over the past 12 months. **LO 1, 2, 3, 5**

DATA*file* monthlysales

Month	Sales	Month	Sales
1	105	7	145
2	135	8	140
3	120	9	100
4	105	10	80
5	90	11	100
6	120	12	110

a. Construct a time series plot. What type of pattern exists in the data?
b. Use $\alpha = 0.3$ to compute the exponential smoothing values for the time series.
c. Use trial and error to find a value of the exponential smoothing coefficient α that results in a relatively small MSE.

DATA*file* futuresindex

15. **Commodity Futures Index.** Twenty weeks of data on the Commodity Futures Index are as follows. **LO 1, 2, 3, 5**

7.35 7.40 7.55 7.56 7.60 7.52 7.52 7.70 7.62 7.55
7.51 7.68 7.85 7.72 7.85 7.73 7.76 7.86 7.81 7.64

a. Construct a time series plot. What type of pattern exists in the data?
b. Use trial and error to find a value of the exponential smoothing coefficient α that results in a relatively small MSE.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

16. **Portfolio Composition.** The following table reports the percentage of stocks in a portfolio for nine quarters. LO 1, 2, 3, 5

DATA*file* portfolio

Quarter	Stock (%)
Year 1, Quarter 1	29.8
Year 1, Quarter 2	31.0
Year 1, Quarter 3	29.9
Year 1, Quarter 4	30.1
Year 2, Quarter 1	32.2
Year 2, Quarter 2	31.5
Year 2, Quarter 3	32.0
Year 2, Quarter 4	31.9
Year 3, Quarter 1	30.0

a. Construct a time series plot. What type of pattern exists in the data?
b. Use trial and error to find a value of the exponential smoothing coefficient α that results in a relatively small MSE.
c. Using the exponential smoothing model you developed in part (b), what is the forecast of the percentage of stocks in a typical portfolio for the second quarter of year 3?

17. **Using Linear Regression for Forecasting with Twenty-Five Time Periods of Data.** Consider the following time series. LO 1, 3, 6

DATA*file* twentyfiveperiods

t	y_t	t	y_t	t	y_t	t	y_t	t	y_t
1	6	6	19	11	28	16	38	21	49
2	11	7	19	12	29	17	39	22	50
3	9	8	22	13	31	18	43	23	52
4	14	9	24	14	34	19	44	24	55
5	15	10	24	15	36	20	45	25	58

a. Construct a time series plot. What type of pattern exists in the data?
b. Use simple linear regression analysis to find the parameters for the line that minimizes MSE for this time series.
c. What is the forecast for $t = 26$?

18. **Using Linear Regression for Forecasting with Twenty-Eight Time Periods of Data.** Consider the following time series. LO 1, 3, 6

DATA*file* twentyeightperiods

t	y_t	t	y_t	t	y_t	t	y_t
1	120	8	82	15	48	22	11
2	110	9	76	16	43	23	5
3	100	10	66	17	39	24	−1
4	96	11	67	18	29	25	−2
5	94	12	61	19	25	26	−6
6	92	13	53	20	24	27	−11
7	88	14	50	21	15	28	−20

a. Construct a time series plot. What type of pattern exists in the data?
b. Use simple linear regression analysis to find the parameters for the line that minimizes MSE for this time series.
c. What is the forecast for $t = 29$?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

19. **University Enrollment.** Because of high tuition costs at state and private universities, enrollments at community colleges have increased in recent years. The following data show the enrollment for Jefferson Community College for the nine most recent years. **LO 1, 3, 5**

DATA*file* jefferson

Year	Period (*t*)	Enrollment (1,000s)
2015	1	6.5
2016	2	8.1
2017	3	8.4
2018	4	10.2
2019	5	12.5
2020	6	13.3
2021	7	13.7
2022	8	17.2
2023	9	18.1

a. Construct a time series plot. What type of pattern exists in the data?
b. Use simple linear regression analysis to find the parameters for the line that minimizes MSE for this time series.
c. What is the forecast for year 10?

20. **Administrative Expenses.** The Seneca Children's Fund (SCF) is a local charity that runs a summer camp for disadvantaged children. The fund's board of directors has been working very hard over recent years to decrease the amount of overhead expenses, a major factor in how charities are rated by independent agencies. The following data show the percentage of the money SCF has raised that was spent on administrative and fund-raising expenses over the past seven years. **LO 1, 3, 6**

DATA*file* seneca

Period (*t*)	Expense (%)
1	13.9
2	12.2
3	10.5
4	10.4
5	11.5
6	10.0
7	8.5

a. Construct a time series plot. What type of pattern exists in the data?
b. Use simple linear regression analysis to find the parameters for the line that minimizes MSE for this time series.
c. Forecast the percentage of administrative expenses for year 8.
d. If SCF can maintain its current trend in reducing administrative expenses, how long will it take for SCF to achieve a level of 5% or less?

21. **Manufacturing Costs.** The president of a small manufacturing firm is concerned about the continual increase in manufacturing costs over the past several years. The following figures provide a time series of the quarterly cost per unit for the firm's leading product over the past eight years. **LO 1, 3, 6**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA file
mfgcosts

Year	Quarter	Cost/Unit ($)	Year	Quarter	Cost/Unit ($)
1	1	20.00	5	1	28.90
	2	17.23		2	32.61
	3	13.39		3	28.66
	4	17.94		4	25.27
2	1	20.29	6	1	25.80
	2	25.61		2	35.71
	3	22.62		3	27.26
	4	29.63		4	27.89
3	1	23.87	7	1	28.50
	2	20.35		2	28.64
	3	26.32		3	32.52
	4	25.50		4	28.24
4	1	20.90	8	1	37.95
	2	28.48		2	28.59
	3	32.65		3	37.07
	4	25.78		4	38.16

a. Construct a time series plot. What type of pattern exists in the data?
b. Use simple linear regression analysis to find the parameters for the line that minimizes MSE for this time series.
c. What is the average cost increase that the firm has been realizing per quarter?
d. Compute an estimate of the cost/unit for the next quarter.

22. **Estimating Seasonal Effects.** Consider the following time series. LO 1, 3, 7

DATA file
seasonaleffects

Quarter	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
1	71	68	62	67	71	64
2	49	41	51	41	46	47
3	58	60	53	78	78	78
4	78	81	72	78	82	80

a. Construct a time series plot. What type of pattern exists in the data? Is there an indication of a seasonal pattern?
b. Use a multiple linear regression model with dummy variables as follows to develop an equation to account for seasonal effects in the data: Qtr1 = 1 if quarter 1, 0 otherwise; Qtr2 = 1 if quarter 2, 0 otherwise; Qtr3 = 1 if quarter 3, 0 otherwise.
c. Compute the quarterly forecasts for the next year.

23. **Estimating Trend and Seasonal Effects.** Consider the following time series data. LO 1, 2, 3, 7

DATA file
trendseasonaleffects

Quarter	Year 1	Year 2	Year 3
1	4	6	7
2	2	3	6
3	3	5	6
4	5	7	8

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Construct a time series plot. What type of pattern exists in the data?
b. Use a multiple regression model with dummy variables as follows to develop an equation to account for seasonal effects in the data: Qtr1 = 1 if quarter 1, 0 otherwise; Qtr2 = 1 if quarter 2, 0 otherwise; Qtr3 = 1 if quarter 3, 0 otherwise.
c. Compute the quarterly forecasts for the next year based on the model you developed in part (b).
d. Use a multiple regression model to develop an equation to account for trend and seasonal effects in the data. Use the dummy variables you developed in part (b) to capture seasonal effects and create a variable t such that $t = 1$ for quarter 1 in year 1, $t = 2$ for quarter 2 in year 1, . . . $t = 12$ for quarter 4 in year 3.
e. Compute the quarterly forecasts for the next year based on the model you developed in part (d).
f. Is the model you developed in part (b) or the model you developed in part (d) more effective? Justify your answer.

24. **Textbook Sales.** The quarterly sales data (number of copies sold) for a college textbook over the past three years are as follows. **LO 1, 2, 3, 7**

DATA*file* textbooksales

Year	1	1	1	1	2	2	2	2	3	3	3	3
Quarter	1	2	3	4	1	2	3	4	1	2	3	4
Sales	1,690	940	2,625	2,500	1,800	900	2,900	2,360	1,850	1,100	2,930	2,615

a. Construct a time series plot. What type of pattern exists in the data?
b. Use a regression model with dummy variables as follows to develop an equation to account for seasonal effects in the data: Qtr1 = 1 if quarter 1, 0 otherwise; Qtr2 = 1 if quarter 2, 0 otherwise; Qtr3 = 1 if quarter 3, 0 otherwise.
c. Based on the model you developed in part (b), compute the quarterly forecasts for the next year.
d. Let $t = 1$ refer to the observation in quarter 1 of year 1; $t = 2$ refer to the observation in quarter 2 of year 1; … ; and $t = 12$ refer to the observation in quarter 4 of year 3. Using the dummy variables defined in part (b) and t, develop an equation to account for seasonal effects and any linear trend in the time series.
e. Based upon the seasonal effects in the data and linear trend, compute the quarterly forecasts for the next year.
f. Is the model you developed in part (b) or the model you developed in part (d) more effective? Justify your answer.

25. **Air Pollution.** Air pollution control specialists in Southern California monitor the amount of ozone, carbon dioxide, and nitrogen dioxide in the air on an hourly basis. The hourly time series data exhibit seasonality, with the levels of pollutants showing patterns that vary over the hours in the day. On July 15, 16, and 17, the following levels of nitrogen dioxide (measured in parts per billion) were observed for the 12 hours from 6:00 a.m. to 6:00 p.m. **LO 1, 2, 3, 7**

DATA*file* pollution

July 15	25	28	35	50	60	60	40	35	30	25	25	20
July 16	28	30	35	48	60	65	50	40	35	25	20	20
July 17	35	42	45	70	72	75	60	45	40	25	25	25

a. Construct a time series plot. What type of pattern exists in the data?
b. Use a multiple linear regression model with dummy variables as follows to develop an equation to account for seasonal effects in the data:

Hour1 = 1 if the reading was made between 6:00 a.m. and 7:00 a.m., 0 otherwise
Hour2 = 1 if the reading was made between 7:00 a.m. and 8:00 a.m., 0 otherwise
⋮
Hour11 = 1 if the reading was made between 4:00 p.m. and 5:00 p.m., 0 otherwise

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note that when the values of the 11 dummy variables are equal to 0, the observation corresponds to the 5:00 p.m. to 6:00 p.m. hour.

c. Using the equation developed in part (b), compute estimates of the levels of nitrogen dioxide for July 18.

d. Let $t = 1$ refer to the observation in hour 1 on July 15; $t = 2$ refer to the observation in hour 2 of July 15; . . . ; and $t = 36$ refer to the observation in hour 12 of July 17. Using the dummy variables defined in part (b) and t_s, develop an equation to account for seasonal effects and any linear trend in the time series.

e. Based on the seasonal effects in the data and linear trend estimated in part (d), compute estimates of the levels of nitrogen dioxide for July 18.

f. Is the model you developed in part (b) or the model you developed in part (d) more effective? Justify your answer.

26. **Sales of Docks and Seawalls.** South Shore Construction builds permanent docks and seawalls along the southern shore of Long Island, New York. Although the firm has been in business for only five years, revenue has increased from $308,000 in the first year of operation to $1,084,000 in the most recent year. The following data show the quarterly sales revenue in thousands of dollars. **LO 1, 2, 3, 7**

DATA*file* southshore

Quarter	Year 1	Year 2	Year 3	Year 4	Year 5
1	20	37	75	92	176
2	100	136	155	202	282
3	175	245	326	384	445
4	13	26	48	82	181

a. Construct a time series plot. What type of pattern exists in the data?

b. Use a multiple regression model with dummy variables as follows to develop an equation to account for seasonal effects in the data: Qtr1 = 1 if quarter 1, 0 otherwise; Qtr2 = 1 if quarter 2, 0 otherwise; Qtr3 = 1 if quarter 3, 0 otherwise.

c. Based on the model you developed in part (b), compute estimates of quarterly sales for year 6.

d. Let Period = 1 refer to the observation in quarter 1 of year 1; Period = 2 refer to the observation in quarter 2 of year 1; . . . ; and Period = 20 refer to the observation in quarter 4 of year 5. Using the dummy variables defined in part (b) and the variable Period, develop an equation to account for seasonal effects and any linear trend in the time series.

e. Based on the seasonal effects in the data and linear trend estimated in part (c), compute estimates of quarterly sales for year 6.

f. Is the model you developed in part (b) or the model you developed in part (d) more effective? Justify your answer.

DATA*file* icecreamsales

27. **Sales of Frozen Treats.** Hogs & Dawgs is an ice cream parlor on the border of north-central Louisiana and southern Arkansas that serves 43 flavors of ice creams, sherbets, frozen yogurts, and sorbets. During the summer Hogs & Dawgs is open from 1:00 p.m. to 10:00 p.m. on Monday through Saturday, and the owner believes that sales change systematically from hour to hour throughout the day. The owner also believes that their sales increase as the outdoor temperature increases. Hourly sales and the outside temperature at the start of each hour for the last week are provided in the file *icecreamsales*. **LO 1, 2, 3, 7**

a. Construct a time series plot of hourly sales and a scatter plot of outdoor temperature and hourly sales. What types of relationships exist in the data?

b. Use a simple regression model with outside temperature as the causal variable to develop an equation to account for the relationship between outside temperature and hourly sales in the data. Based on this model, compute an estimate of hourly sales for today from 2:00 p.m. to 3:00 p.m. if the temperature at 2:00 p.m. is 93°F.

c. Use a multiple linear regression model with the causal variable outside temperature and dummy variables as follows to develop an equation to account for both

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

seasonal effects and the relationship between outside temperature and hourly sales in the data:

Hour1 = 1 if the sales were recorded between 1:00 p.m. and 2:00 p.m., 0 otherwise
Hour2 = 1 if the sales were recorded between 2:00 p.m. and 3:00 p.m., 0 otherwise
⋮
Hour8 = 1 if the sales were recorded between 8:00 p.m. and 9:00 p.m., 0 otherwise

Note that when the values of the eight dummy variables are equal to 0, the observation corresponds to the 9:00-to-10:00-p.m. hour.

Based on this model, compute an estimate of hourly sales for today from 2:00 p.m. to 3:00 p.m. if the temperature at 2:00 p.m. is 93°F.

d. Is the model you developed in part (b) or the model you developed in part (c) more effective? Justify your answer.

DATA *file* gasstation

28. **Gasoline Sales and Price.** Donna Nickles manages a gasoline station on the corner of Bristol Avenue and Harpst Street in Arcata, California. Donna's station is a franchise, and the parent company calls the station every day at midnight to give Donna the prices for various grades of gasoline for the upcoming day. Over the past eight weeks Donna has recorded the price and sales (in gallons) of regular-grade gasoline at the station as well as the price of regular-grade gasoline charged by the competitor across the street. Donna is curious about the sensitivity of their sales to the price of regular gasoline they charge and the price of regular gasoline charged by the competitor across the street. Donna also wonders whether their sales differ systematically by day of the week and whether their station has experienced a trend in sales over the past eight weeks. The data collected by Donna for each day of the past eight weeks are provided in the file *gasstation.* **LO 1, 2, 3, 7**

a. Construct a time series plot of daily sales, a scatter plot of the price Donna charges for a gallon of regular gasoline and daily sales at Donna's station, and a scatter plot of the price Donna's competitor charges for a gallon of regular gasoline and daily sales at Donna's station. What types of relationships exist in the data?

b. Use a multiple regression model with the price Donna charges for a gallon of regular gasoline and the price Donna's competitor charges for a gallon of regular gasoline as causal variables to develop an equation to account for the relationships between these prices and Donna's daily sales in the data. Based on this model, compute an estimate of sales for a day on which Donna is charging $3.50 for a gallon of regular gasoline and Donna's competitor is charging $3.45 for a gallon of regular gasoline.

c. Use a multiple linear regression model with the trend and dummy variables as follows to develop an equation to account for both trend and seasonal effects in the data:

Monday = 1 if the sales were recorded on a Monday, 0 otherwise
Tuesday = 1 if the sales were recorded on a Tuesday, 0 otherwise
⋮
Saturday = 1 if the sales were recorded on a Saturday, 0 otherwise

Note that when the values of the six dummy variables are equal to 0, the observation corresponds to Sunday.

Based on this model, compute an estimate of sales for Tuesday of the first week after Donna collected the data.

d. Use a multiple regression model with the price Donna charges for a gallon of regular gasoline and the price Donna's competitor charges for a gallon of regular gasoline as causal variables and the trend and dummy variables from part (c) to create an equation to account for the relationships between these prices and daily sales as well as the trend and seasonal effects in the data. Based on this model, compute an estimate of sales for Tuesday of the first week after Donna collected the data a day if Donna is charging $3.50 for a gallon of regular gasoline and Donna's competitor is charging $3.45 for a gallon of regular gasoline.

e. Which of the three models you developed in parts (b), (c), and (d) is most effective? Justify your answer.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Case Problem 1: Forecasting Food and Beverage Sales

The Vintage Restaurant, on Captiva Island near Fort Myers, Florida, is owned and operated by Karen Payne. The restaurant just completed its third year of operation. During those three years, Karen sought to establish a reputation for the restaurant as a high-quality dining establishment that specializes in fresh seafood. Through the efforts of Karen and their staff, their restaurant has become one of the best and fastest-growing restaurants on the Island.

To better plan for future growth of the restaurant, Karen needs to develop a system that will enable them to forecast food and beverage sales by month for up to one year in advance. The following table shows the value of food and beverage sales ($1,000s) for the first three years of operation. **LO 1, 3, 7**

DATA*file*
vintage

Month	First Year	Second Year	Third Year
January	242	263	282
February	235	238	255
March	232	247	265
April	178	193	205
May	184	193	210
June	140	149	160
July	145	157	166
August	152	161	174
September	110	122	126
October	130	130	148
November	152	167	173
December	206	230	235

Managerial Report

Perform an analysis of the sales data for the Vintage Restaurant. Prepare a report for Karen that summarizes your findings, forecasts, and recommendations. Include the following:

1. A time series plot. Comment on the underlying pattern in the time series.
2. Using the dummy variable approach, forecast sales for January through December of the fourth year. How would you explain this model to Karen?

Assume that January sales for the fourth year turn out to be $295,000. What was your forecast error? If this error is large, Karen may be puzzled about the difference between your forecast and the actual sales value. What can you do to resolve Karen's uncertainty about the forecasting procedure?

Case Problem 2: Forecasting Lost Sales

The Carlson Department Store suffered heavy damage when a hurricane struck on August 31. The store was closed for four months (September through December), and Carlson is now involved in a dispute with its insurance company about the amount of lost sales during the time the store was closed. Two key issues must be resolved: (1) the amount of sales Carlson would have made if the hurricane had not struck and (2) whether Carlson is entitled to any compensation for excess sales due to increased business activity after the storm. More than $8 billion in federal disaster relief and insurance money came into the county, resulting in increased sales at department stores and numerous other businesses.

The following two tables give (1) Carlson's sales data for the 48 months preceding the storm and (2) the total sales for the 48 months preceding the storm for all department stores in the county, as well as the total sales in the county for the four months the Carlson Department Store was closed. Carlson's managers asked you to analyze these data and develop estimates of the lost sales at the Carlson Department Store for the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

months of September through December. They also asked you to determine whether a case can be made for excess storm-related sales during the same period. If such a case can be made, Carlson is entitled to compensation for excess sales it would have earned in addition to ordinary sales. **LO 1, 7**

Managerial Report

Prepare a report for the managers of the Carlson Department Store that summarizes your findings, forecasts, and recommendations. Include the following:

1. An estimate of sales for Carlson Department Store had there been no hurricane.
2. An estimate of countywide department store sales had there been no hurricane.
3. An estimate of lost sales for the Carlson Department Store for September through December.

In addition, use the countywide actual department stores sales for September through December and the estimate in part (2) to make a case for or against excess storm-related sales.

DATA *file*
carlsonsales

Sales for Carlson Department Store ($ millions)

Month	Year 1	Year 2	Year 3	Year 4	Year 5
January		1.45	2.31	2.31	2.56
February		1.80	1.89	1.99	2.28
March		2.03	2.02	2.42	2.69
April		1.99	2.23	2.45	2.48
May		2.32	2.39	2.57	2.73
June		2.20	2.14	2.42	2.37
July		2.13	2.27	2.40	2.31
August		2.43	2.21	2.50	2.23
September	1.71	1.90	1.89	2.09	
October	1.90	2.13	2.29	2.54	
November	2.74	2.56	2.83	2.97	
December	4.20	4.16	4.04	4.35	

DATA *file*
countysales

Department Store Sales for the County ($ millions)

Month	Year 1	Year 2	Year 3	Year 4	Year 5
January		46.80	46.80	43.80	48.00
February		48.00	48.60	45.60	51.60
March		60.00	59.40	57.60	57.60
April		57.60	58.20	53.40	58.20
May		61.80	60.60	56.40	60.00
June		58.20	55.20	52.80	57.00
July		56.40	51.00	54.00	57.60
August		63.00	58.80	60.60	61.80
September	55.80	57.60	49.80	47.40	69.00
October	56.40	53.40	54.60	54.60	75.00
November	71.40	71.40	65.40	67.80	85.20
December	117.60	114.00	102.00	100.20	121.80

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 9 Appendix

Appendix 9.1 Using the Excel Forecast Sheet

Excel features a tool called Forecast Sheet which can automatically produce forecasts using the Holt–Winters additive seasonal smoothing model. The Holt–Winters model is an exponential smoothing approach to estimating additive linear trend and seasonal effects. It also generates a variety of other outputs that are useful in assessing the accuracy of the forecast model it produces.

Forecast Sheet was introduced in Excel 2016; it is not available in prior versions of Excel or in Excel Online.

We will demonstrate Forecast Sheet on the four years of quarterly smartphone sales that are provided in Table 9.6. A review of the time series plot of these data in Figure 9.6 provides clear evidence of an increasing linear trend and a seasonal pattern (sales are consistently lowest in the second quarter of each year and highest in quarters 3 and 4). We concluded in Section 9.4 that we need to use a forecasting method that is capable of dealing with both trend and seasonality when developing a forecasting model for this time series, and so it is appropriate to use Forecast Sheet to produce forecasts for these data.

Excel refers to the forecasting approach used by Forecast Sheet as the AAA exponential smoothing (ETS) algorithm, where AAA stands for additive error, additive trend, and additive seasonality.

We begin by putting the data into the format required by Forecast Sheet. The time series data must be collected on a consistent interval (i.e., annually, quarterly, monthly, and so on), and the spreadsheet must include two data series in contiguous columns or rows that include

- a series with the dates or periods in the time series
- a series with corresponding time series values

DATA*file*
smartphonesales

First, open the file *smartphonesales*, then insert a column between column B ("Quarter") and Column C ("Sales (1000s)"). Enter *Period* into cell C1; this will be the heading for the column of values that will represent the periods in our data. Next enter *1* in cell C2, *2* in cell C3, *3* in cell C4, and so on, ending with *16* in Cell C17, as shown in Figure 9.19.

Now that the data are properly formatted for Forecast Sheet, the following steps can be used to produce forecasts for the next four quarters (periods 17 through 20) with Forecast Sheet:

Step 1. Highlight cells C1:D17 (the data in column C of this highlighted section is what Forecast Sheet refers to as the **Timeline Range** and the data in column D is the **Values Range**)
Step 2. Click the **Data** tab in the Ribbon
Step 3. Click **Forecast Sheet** in the **Forecast** group
Step 4. When the **Create Forecast Worksheet** dialog box appears (Figure 9.20):
Select **20** for **Forecast End**
Click **Options** to expand the **Create Forecast Worksheet** dialog box and show the options (Figure 9.20)
Select **16** for **Forecast Start**
Select **95%** for **Confidence Interval**
Under **Seasonality**, click on **Set Manually** and select **4**
Select the checkbox for **Include forecast statistics**
Click **Create**

Forecast Sheet requires that the period selected for **Forecast Start** *is one of the periods of the original time series.*

The results of Forecast Sheet will be output to a new worksheet as shown in Figure 9.21. The output of Forecast Sheet includes the following:

- The period for each of the 16 time series observations and the forecasted time periods in column A
- The actual time series data for periods 1 to 16 in column B

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.19 Smartphone Data Reformatted for Forecast Sheet

	A	B	C	D
1	Year	Quarter	Period	Sales (1000s)
2	1	1	1	4.8
3	1	2	2	4.1
4	1	3	3	6.0
5	1	4	4	6.5
6	2	1	5	5.8
7	2	2	6	5.2
8	2	3	7	6.8
9	2	4	8	7.4
10	3	1	9	6.0
11	3	2	10	5.6
12	3	3	11	7.5
13	3	4	12	7.8
14	4	1	13	6.3
15	4	2	14	5.9
16	4	3	15	8.0
17	4	4	16	8.4

- The forecasts for periods 16 to 20 in column C
- The lower confidence bounds for the forecasts for periods 16 to 20 in column D
- The upper confidence bounds for the forecasts for periods 16 to 20 in column E
- A line graph of the time series, forecast values, and forecast interval
- The values of the three parameters (alpha, beta, and gamma) used in the Holt–Winters additive seasonal smoothing model in cells H2:H4 (these values are determined by an algorithm in Forecast Sheet)
- Measures of forecast accuracy in cells H5:H8, including
  - the MASE, or mean absolute scaled error, in cell H5. MASE is defined as:

$$\text{MASE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|e_t|}{\frac{1}{n-1}\sum_{t=2}^{n}|y_t - y_{t-1}|}$$

MASE compares the forecast error, e_t, to a naïve forecast error given by $|y_t - y_{t-1}|$. If MASE > 1, then the forecast is considered inferior to a naïve forecast; if MASE < 1, the forecast is considered superior to a naïve forecast.

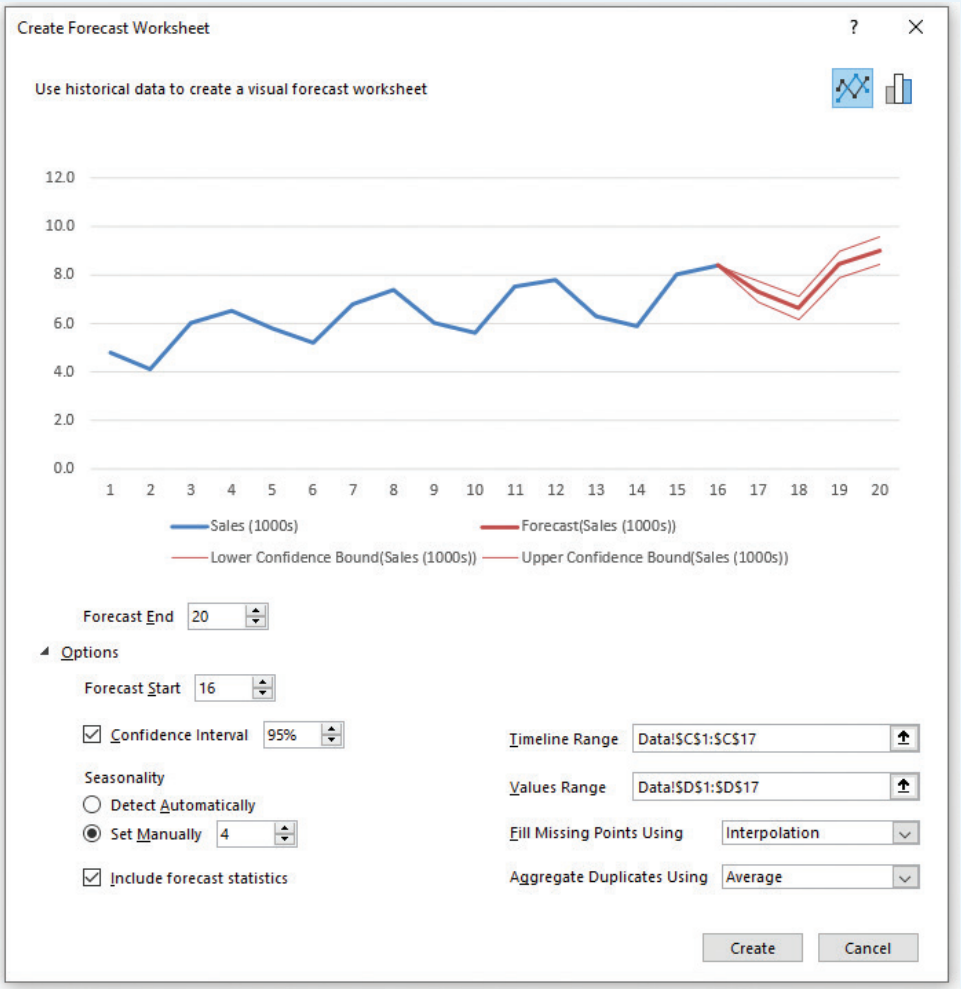
  - the SMAPE, or symmetric mean absolute percentage error, in cell H6. SMAPE is defined as:

$$\text{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|e_t|}{(|y_t| + |\hat{y}_t|)/2}$$

SMAPE is similar to mean absolute percentage error (MAPE), discussed in Section 9.2; both SMAPE and MAPE measure forecast error relative to actual values.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.20 Create Forecast Worksheet Dialog Box with Options Open for Quarterly Smartphone Sales



- the MAE, or mean absolute error, (as defined in equation (9.3)) in cell H7
- the RMSE, or root mean squared error, (which is the square root of the MSE, defined in equation (9.4)) in cell H8

Figures 9.22 and 9.23 display the formula view of portions of the worksheet that Forecast Sheet generated based on the smartphone quarterly sales data. For example, in cell C18, the forecast value generated for Period 17 smartphone sales is determined by the formula:

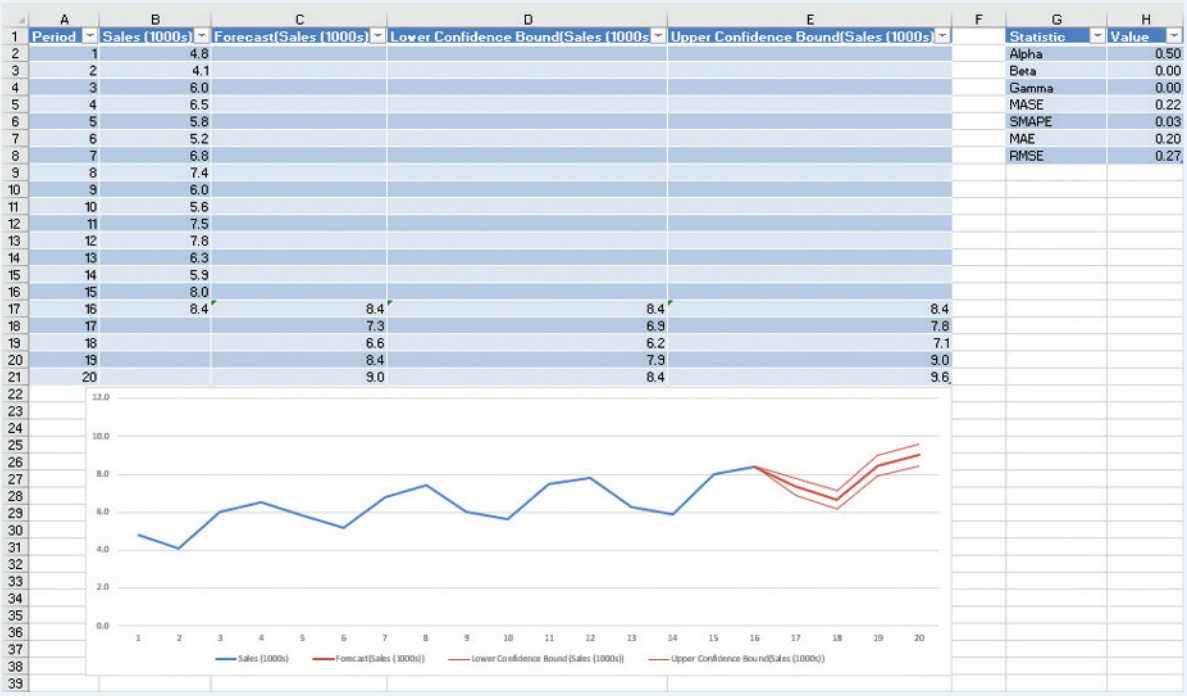
=FORECAST.ETS(A18, B2:B17, A2:A17, 4, 1)

There is a sixth (optional) argument of the FORECAST.ETS function that addresses how to aggregate multiple observations for the same time period. Choices include AVERAGE, SUM, COUNT, COUNTA, MIN, MAX, and MEDIAN.

The first argument in this function specifies the period to be forecasted. The second argument specifies the time series data upon which the forecast is based. The third argument lists the timeline associated with the time series values. The fourth (optional) argument addresses seasonality, and the value of 4 indicates the length of the seasonal pattern. A value of 1 for the fourth argument of the FORECAST.ETS function means that Excel detects the seasonality in the data automatically. A value of 0 means that there is no seasonality in the data. The fifth (optional) argument addresses missing data, and a value of 1 means that any missing observations will be approximated as the average of

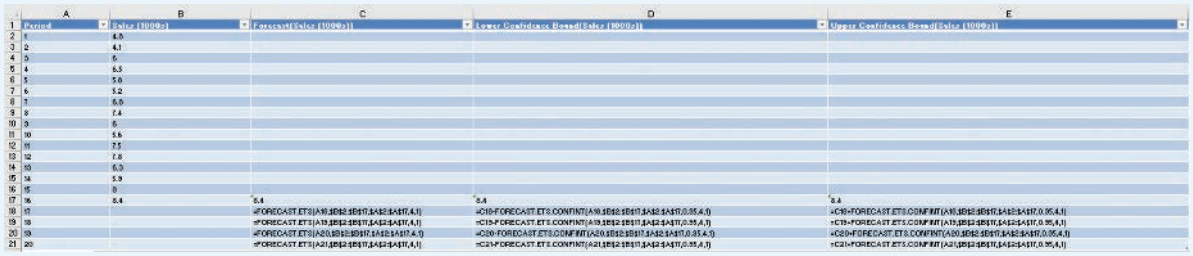
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.21 Forecast Sheet Results for Quarterly Smartphone Sales



	A	B	C	D	E	F	G	H
1	Period	Sales (1000s)	Forecast(Sales (1000s)	Lower Confidence Bound(Sales (1000s	Upper Confidence Bound(Sales (1000s)		Statistic	Value
2	1	4.8					Alpha	0.50
3	2	4.1					Beta	0.00
4	3	6.0					Gamma	0.00
5	4	6.5					MASE	0.22
6	5	5.8					SMAPE	0.03
7	6	5.2					MAE	0.20
8	7	6.8					RMSE	0.27
9	8	7.4						
10	9	6.0						
11	10	5.6						
12	11	7.5						
13	12	7.8						
14	13	6.3						
15	14	5.9						
16	15	8.0						
17	16	8.4	8.4	8.4	8.4			
18	17		7.3	6.9	7.8			
19	18		6.6	6.2	7.1			
20	19		8.4	7.9	9.0			
21	20		9.0	8.4	9.6			

Figure 9.22 Formula View of Forecast Sheet Results for Quarterly Smartphone Sales



	A	B	C	D	E
1	Period	Sales (1000s)	Forecast(Sales (1000s))	Lower Confidence Bound(Sales (1000s))	Upper Confidence Bound(Sales (1000s))
2	1	4.8			
3	2	4.1			
4	3	6			
5	4	6.5			
6	5	5.8			
7	6	5.2			
8	7	6.8			
9	8	7.4			
10	9	6			
11	10	5.6			
12	11	7.5			
13	12	7.8			
14	13	6.3			
15	14	5.9			
16	15	8			
17	16	8.4	8.4	8.4	8.4
18	17		=FORECAST.ETS(A18,$B$2:$B$17,$A$2:$A$17,4,1)	=C18-FORECAST.ETS.CONFINT(A18,$B$2:$B$17,$A$2:$A$17,0.95,4,1)	=C18+FORECAST.ETS.CONFINT(A18,$B$2:$B$17,$A$2:$A$17,0.95,4,1)
19	18		=FORECAST.ETS(A19,$B$2:$B$17,$A$2:$A$17,4,1)	=C19-FORECAST.ETS.CONFINT(A19,$B$2:$B$17,$A$2:$A$17,0.95,4,1)	=C19+FORECAST.ETS.CONFINT(A19,$B$2:$B$17,$A$2:$A$17,0.95,4,1)
20	19		=FORECAST.ETS(A20,$B$2:$B$17,$A$2:$A$17,4,1)	=C20-FORECAST.ETS.CONFINT(A20,$B$2:$B$17,$A$2:$A$17,0.95,4,1)	=C20+FORECAST.ETS.CONFINT(A20,$B$2:$B$17,$A$2:$A$17,0.95,4,1)
21	20		=FORECAST.ETS(A21,$B$2:$B$17,$A$2:$A$17,4,1)	=C21-FORECAST.ETS.CONFINT(A21,$B$2:$B$17,$A$2:$A$17,0.95,4,1)	=C21+FORECAST.ETS.CONFINT(A21,$B$2:$B$17,$A$2:$A$17,0.95,4,1)

the neighboring observations. A value of 0 for the fifth argument of the FORECAST.ETS function means that Excel will treat any missing observations as zeros. These data have no missing observations, so the value of this fifth argument does not matter.

Cell D18 contains the lower confidence bound for the forecast of the Period 17 smartphone sales. This lower confidence bound is determined by the formula:

$$=C18 - \text{FORECAST.ETS.CONFINT}(A18, B2:B17, A2:A17, 0.95, 4, 1)$$

Similarly, cell E18 contains the upper confidence bound for the forecast of the Period 17 smartphone sales. This upper confidence bound is determined by the formula:

$$=C18 + \text{FORECAST.ETS.CONFINT}(A18, B2:B17, A2:A17, 0.95, 4, 1)$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 9.23 Excel Formulas for Smartphone Forecast Statistics

	F	G	H
1		Statistic	Value
2		Alpha	=FORECAST.ETS.STAT($B$2:$B$17,$A$2:$A$17,1,4,1)
3		Beta	=FORECAST.ETS.STAT($B$2:$B$17,$A$2:$A$17,2,4,1)
4		Gamma	=FORECAST.ETS.STAT($B$2:$B$17,$A$2:$A$17,3,4,1)
5		MASE	=FORECAST.ETS.STAT($B$2:$B$17,$A$2:$A$17,4,4,1)
6		SMAPE	=FORECAST.ETS.STAT($B$2:$B$17,$A$2:$A$17,5,4,1)
7		MAE	=FORECAST.ETS.STAT($B$2:$B$17,$A$2:$A$17,6,4,1)
8		RMSE	=FORECAST.ETS.STAT($B$2:$B$17,$A$2:$A$17,7,4,1)

Many of the arguments for the FORECAST.ETS.CONFINT function are the same as the arguments for the FORECAST.ETS function. The first argument in the FORECAST.ETS.CONFINT function specifies the period to be forecasted. The second argument specifies the time series data upon which the forecast is based. The third argument lists the timeline associated with the time series values. The fourth (optional) argument specifies confidence level associated with the calculated confidence interval. The fifth (optional) argument addresses seasonality, and the value of 4 indicates the length of the seasonal pattern. The sixth (optional) argument addresses missing data, and a value of 1 means that any missing observations will be approximated as the average of the neighboring observations; this data had no missing observations, so the value of this argument does not matter.

Cells H2:H8 of Figure 9.23 list the Excel formulas used to compute the respective statistics for the smartphone sales forecasts. These formulas are:

- Alpha

  =FORECAST.ETS.STAT(B2:B17, A2:A17, 1, 4, 1)

- Beta

  =FORECAST.ETS.STAT(B2:B17, A2:A17, 2, 4, 1)

- Gamma

  =FORECAST.ETS.STAT(B2:B17, A2:A17, 3, 4, 1)

- MASE

  =FORECAST.ETS.STAT(B2:B17, A2:A17, 4, 4, 1)

- SMAPE

  =FORECAST.ETS.STAT(B2:B17, A2:A17, 5, 4, 1)

- MAE

  =FORECAST.ETS.STAT(B2:B17, A2:A17, 6, 4, 1)

- RMSE

  =FORECAST.ETS.STAT(B2:B17, A2:A17, 7, 4, 1)

Many of the arguments for the FORECAST.ETS.STAT function are the same as the arguments for the FORECAST.ETS function. The first argument in the FORECAST.ETS.STAT function the time series data upon which the forecast is based. The second argument lists the timeline associated with the time series values. The third argument specifies the statistic or parameter type; for example, a value of 4 corresponds to MASE statistic. The fourth (optional) argument addresses seasonality, and the value of 4 indicates the length of the seasonal pattern. The fifth

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(optional) argument addresses missing data, and a value of 1 means that any missing observations will be approximated as the average of the neighboring observations; this data had no missing observations, so the value of this argument does not matter.

We conclude this appendix with a few comments on the functionality of Forecast Sheet. Forecast Sheet includes an algorithm for automatically finding the number of time periods over which the seasonal pattern recurs. To use this algorithm, select the option for **Detect Automatically** under **Seasonality** in the **Create Forecast Worksheet** dialog box before clicking **Create**. We suggest using this feature only to confirm a suspected seasonal pattern as using this feature to find a seasonal effect may lead to identification of a spurious pattern that does not actually reflect seasonality. This would result in a model that is overfit on the observed time series data and would likely produce very inaccurate forecasts. A forecast model with seasonality should only be fit when the modeler has reason to suspect a specific seasonal pattern.

The **Forecast Start** parameter in the **Create Forecast Worksheet** dialog box controls both the first period to be forecasted and the last period to be used to generate the forecast model. If we had selected 15 for Forecast Start, we would have generated a forecast model for the smartphone monthly sales data based on only the first 15 periods of data in the original time series.

Forecast Sheet can accommodate multiple observations for a single period of the time series. The **Aggregate Duplicates Using** option in the **Create Forecast Worksheet** dialog box allows the user to select from several ways to deal with this issue.

Forecast Sheet allows for up to 30% of the values for the time series variable to be missing. In the smartphone quarterly sales data, the value of sales for up to 30% of the 16 periods (or 4 periods) could be missing and Forecast Sheet will still produce forecasts. The **Fill Missing Points Using** option in the **Create Forecast Worksheet** dialog box allows the user to select whether the missing values will be replaced with zero or with the result of linearly interpolating existing values in the time series.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 10

Predictive Data Mining: Regression Tasks

Contents

Analytics in Action: Orbitz

10.1 Regression Performance Measures

10.2 Data Sampling, Preparation, and Partitioning
- Static Holdout Method
- *k*-Fold Cross-Validation

10.3 *k*-Nearest Neighbors Regression

10.4 Regression Trees
- Constructing a Regression Tree
- Generating Predictions with a Regression Tree
- Ensemble Methods

10.5 Neural Network Regression
- Structure of a Neural Network
- How a Neural Network Learns

10.6 Feature Selection
- Wrapper Methods
- Filter Methods
- Embedded Methods

Summary 558

Glossary 558

Problems 560

Available in the Cengage eBook:
- Appendix: *k*-Nearest Neighbors Regression with R
- Appendix: Individual Regression Trees with R
- Appendix: Random Forests of Regression Trees with R
- Appendix: Neural Network Regression with R
- Appendix: Regularized Linear Regression with R
- Appendix: *k*-Nearest Neighbors Regression with Orange
- Appendix: Individual Regression Trees with Orange
- Appendix: Random Forests of Regression Trees with Orange
- Appendix: Neural Network Regression with Orange
- Appendix: Regularized Linear Regression with Orange

Learning Objectives

After completing this chapter, you will be able to

LO 1 Compute and interpret performance measures for prediction methods on regression tasks.

LO 2 Evaluate the out-of-sample predictive performance of a model with a static holdout and *k*-fold cross-validation experiment.

LO 3 Describe and apply *k*-nearest neighbors to make predictions for a regression task.

LO 4 Describe and apply regression trees to make predictions of a quantitative outcome variable.

LO 5 Describe and apply ensemble methods to make predictions for a regression task.

LO 6 Describe and apply neural networks to make predictions for a regression task.

LO 7 Describe various ways of selecting features for predictive models for a regression task.

LO 8 Describe the use of regularization to avoid the overfitting of a predictive model.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

Orbitz.com*

Although they might not see their customers face to face, online retailers are getting to know their patrons in order to tailor the offerings on their virtual shelves. By mining web-browsing data collected in "cookies"—files that web sites use to track people's web-browsing behavior, online retailers identify trends that can potentially be used to improve customer satisfaction and boost online sales.

Orbitz.com is an online travel fare aggregator owned by Expedia Group where users can book flights, hotels, car rentals, cruises, and other travel activities. By tracking its customers' online activities, Orbitz.com discovered that people who use Mac computers spend as much as 30% more per night on hotels. Orbitz.com's analytics team has uncovered other factors that affect purchase behavior, including how the shopper arrived at the Orbitz.com site (Did the user visit Orbitz.com directly or was the user referred from another site?), previous booking history on Orbitz.com, and the shopper's geographic location. Orbitz.com can act on this and other information gleaned from mining the vast amount of data on its customers to differentiate the recommendations for hotels, car rentals, flight bookings, etc.

*"On Orbitz, Mac Users Steered to Pricier Hotels," *Wall Street Journal* (2012, June 26).

Organizations are collecting an increasing amount of data, and one of the most pressing tasks is converting these data into actionable insights. A common challenge is to leverage these data to extract information on patterns and trends that can be used to make predictions about the future.

In this chapter, we focus on data mining methods for **regression**, the task of predicting a quantitative outcome using a set of input variables, or **features**. Linear regression is a well-known statistical method for regression tasks in which observations of a quantitative outcome (the dependent variable, y) and one or more corresponding features (the independent variables $x_1, x_2, \ldots, x_q$) are used to create an equation for predicting y values. In this chapter, we introduce other predictive data mining methods for addressing a regression task. Predictive data mining methods are also referred to as **supervised learning** because the known values of the outcome variable "supervise" or guide the process of learning how to use the input features to predict future values of the outcome variable.

In Chapter 8, we describe linear regression modeling in more detail.

In Chapter 11, we discuss data mining methods for addressing the task of predicting a variable when it is categorical instead of quantitative.

The predictive data mining process comprises the following steps:

1. *Data sampling*. Extract a sample of data that is relevant to the business problem under consideration.
2. *Data preparation*. Process the data to put it in a form suitable for formal modeling. This step includes addressing missing and erroneous data, reducing the number of variables, and defining new variables. Data exploration is an important part of this step and may involve the use of descriptive statistics, data visualization, and clustering to better understand the relationships supported by the data.
3. *Data partitioning*. Divide the sample data for the training, validation, and testing of the data mining algorithm performance.
4. *Model construction and assessment*. Construct a set of supervised learning models (e.g., linear regression, k-nearest neighbors, tree-based methods, neural networks) on the training data and make out-of-sample predictions on the validation set to compare their performance. Select a final model and make predictions on the test set as a final appraisal of the selected model's performance.

Chapters 2 and 3 discuss descriptive statistics and data-visualization techniques. Chapter 4 describes the data preparation process. Chapter 6 discusses dimension reduction and clustering techniques that can be used to redefine variables.

10.1 Regression Performance Measures

Palanquin Auto is a national pre-owned car retailer that offers an online marketplace for customers interested in purchasing a pre-owned automobile. In an effort to reduce price negotiation, Palanquin continually collects data on automobile sales in an effort to set fair, competitive prices. Palanquin collects data on recent pre-owned car sales including sales price, vehicle age, vehicle manufacturer, vehicle model, vehicle condition, number of

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

engine cylinders, fuel type, odometer reading, transmission type, wheel-drive, paint color, and geographical coordinates of where the vehicle was sold. Palanquin would like to construct and assess a set of candidate prediction methods for the regression task of predicting the sales price of a pre-owned car.

In Chapter 9, we discuss a variety of measures for assessing accuracy of forecasts of quantitative outcomes.

We begin by reviewing two common ways to measure performance when predicting a quantitative outcome variable. We define the **prediction error** of an observation as the difference between the actual value of an observation and the predicted value of an observation, $y_i - \hat{y}_i$. The **mean absolute error (MAE)** is the average of the absolute values of the prediction errors over a specified set of n observations:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

The MAE calculation considers the absolute value of the prediction errors to avoid cases of under-prediction ($y_i - \hat{y}_i > 0$) from cancelling out cases of over-prediction ($y_i - \hat{y}_i < 0$).

The **root mean squared error (RMSE)** is the square root of the average of the squared prediction errors over a specified set of n observations:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

The RMSE calculation uses the squared prediction errors to avoid cases of under-prediction ($y_i - \hat{y}_i > 0$) from canceling out cases of over-prediction ($y_i - \hat{y}_i < 0$). The term under the square root in the RMSE calculation is called the *mean squared error* (MSE) and it corresponds to the average of the squared prediction errors. To obtain a measure in the same units as the outcome variable, the square root of the mean squared error is applied.

In many cases, when comparing prediction methods, MAE and RMSE will both favor the same prediction method. However, this is not always the case. Consider two methods of predicting eight pre-owned car prices for Palanquin in Table 10.1. The MAE of prediction method 1 is smaller than the MAE of prediction method 2. On average, the absolute value of the prediction error of method 1 is smaller than the absolute value of the prediction error of method 2. However, the RMSE of prediction method 1 is larger than the RMSE of prediction

Table 10.1 Comparing MAE to RMSE for Eight Pre-owned Car Price Predictions

Car	Actual Price ($)	Method 1 Predicted Price ($)	Method 2 Predicted Price ($)	Method 1 Error	Method 2 Error
1	14,000	14,741	12,124	−741	1,876
2	15,000	15,313	24,984	−313	−9,984
3	22,950	15,771	33,147	7,179	−10,197
4	900	802	7,721	98	−6,821
5	20,950	18,895	24,789	2,055	−3,839
6	2,500	7,362	1,037	−4,862	1,463
7	7,300	14,568	19,545	−7,268	−12,245
8	45,000	21,141	49,878	23,859	−4,878
			MAE:	5,797	6,413
			RMSE:	9,368	7,452

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

method 2. This occurs because method 1 has a very large error in predicting the price of car 8 that, when squared in the RMSE calculation, heavily penalizes method 1. To see the discrepancy in how the prediction error for car 8 is assessed differently by MAE and RMSE, observe that the absolute error for Method 1 on car 8 is 23,859 and the absolute error for method 2 on car 7 is 4,878, resulting in a difference in absolute accuracy of 18,981. However, the squared error for Method 1 on car 8 is 569,251,881 and the squared error for Method 1 on car 8 is 23,794,884. The square root of this squared difference is 23,355. Thus, RMSE penalizes Method 1 more than Method 2 because Method 1 results in a larger error for car 8.

10.2 Data Sampling, Preparation, and Partitioning

Upon identifying a business problem, data on relevant variables must be obtained for analysis. Although access to large amounts of data offers the potential to unlock insight and improve decision making, it comes with the risk of being overwhelmed. Data repositories with millions of observations over hundreds of measured variables are now common. If the volume of relevant data is extremely large, it may be unnecessary (and computationally burdensome) to use all the data in order to perform a detailed analysis. When dealing with large volumes of data (with hundreds of thousands or millions of observations), a common practice is to extract a representative sample (with thousands or tens of thousands of observations) for analysis. A sample is representative if the analyst can make the same conclusions from it as from the entire population of data.

There are no definite rules to determine the size of a sample. The sample of data must be large enough to contain significant information, yet small enough to process quickly. If the sample is too small, relationships in the data may be missed or spurious relationships may be suggested. Because data mining algorithms typically are more effective given more data, perhaps the best advice is to use enough data to eliminate any doubt about whether the sample size is sufficient.

Once a representative data sample has been prepared for analysis, it must be partitioned to appropriately evaluate the performance of predictive data mining models. In contrast to applications with a relative scarcity of data in which statistical inference is leveraged to make conclusions about a population, data mining applications deal with an abundance of data. Increasing volumes of data simplifies the process of assessing model performance, but can also lead to **overfitting**. For a regression task, overfitting occurs when the analyst builds a model with a small RMSE or MAE for the sample of data on which the model was constructed, but the RMSE or MAE of the model is much larger for its predictions of observations that were not part of its training. We can use the abundance of data to guard against the potential for overfitting by splitting the data set into different subsets for (1) the training (or construction) of candidate models, (2) the validation (or performance comparison) of candidate models, and (3) the testing of future performance of a selected model.

In the following two sections, we describe two different methods for splitting the data into sets for the training and validation of prediction models: static holdout and *k*-fold cross-validation. With the static holdout method, the analyst partitions the data into static training and validation sets. These are called "static" training and validation sets because the data sets are only constructed once. In contrast, as we describe later, the *k*-fold cross-validation procedure iteratively rotates the data to create *k* training and validation sets.

Static Holdout Method

The most straightforward way to partition data to combat overfitting is to split the sample data into three static data sets: the training set, the validation set, and the test set. The **training set** consists of the data used to build the candidate models. For example, a training set may be used to estimate the slope coefficients for a collection of candidate linear regression models that differ based on the variables included in the models. However, since the training set consists of the data used to build the models, it cannot be used to reliably identify the best model for predicting data outside the training set. Therefore, we use models constructed on the training set to make predictions on the **validation set**, which is a set of observations that have not been used in model training. To compare regression models constructed on the training data, their predictive performance is measured using metrics such as RMSE and MAE on the observations in validation set.

We discuss linear regression models in Chapter 8.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To demonstrate the static holdout method, let us consider the data in Table 10.2. Table 10.2 contains the age (in years) and sales price (in $) of 12 pre-owned cars collected by Palanquin Auto. Suppose Palanquin Auto is interested in constructing and assessing linear regression models that use the age of a car to predict its sales price. Let a static training set consist of the eight observations corresponding to cars 1-8 and a static validation set consist of the four observations corresponding to cars 9–12.

Figure 10.1 illustrates two candidate linear regression models constructed on the static training set consisting of cars 1–8, which are represented by the blue markers. Based on these eight observations, an estimated simple linear regression model is

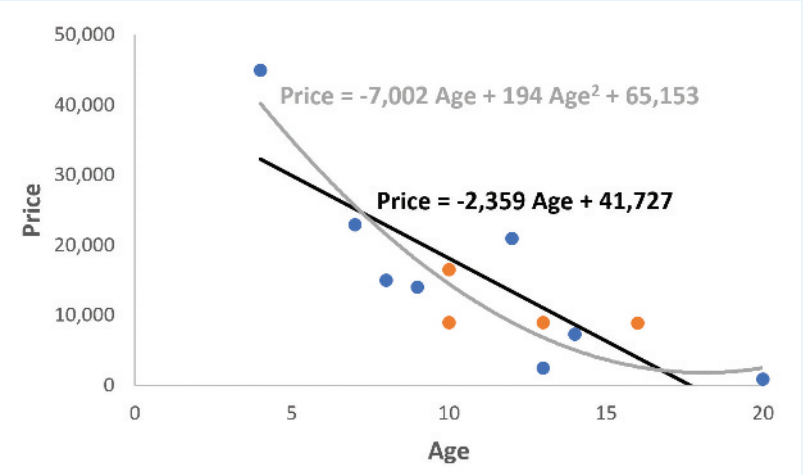
$$\text{Price} = -2{,}359\ \text{Age} + 41{,}727$$

DATA *file*
preowned12.csv

Table 10.2 Static Holdout Method for 12 Observations

Car	Age (years)	Price ($)	
1	9	14,000	
2	8	15,000	
3	7	22,950	
4	20	900	Training Set
5	12	20,950	
6	13	2,500	
7	14	7,300	
8	4	45,000	
9	10	8,950	
10	16	8,888	Validation Set
11	10	16,500	
12	13	9,000	

Figure 10.1 Candidate Linear Regression Models Constructed on Static Training Set (Blue) and Assessed on Static Validation Set (Orange)



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The predictions of this simple linear regression model are represented by the black line in Figure 10.1.

Based on these eight observations, an estimated multiple linear regression model is

$$\text{Price} = -7{,}002\,\text{Age} + 194\,\text{Age}^2 + 65{,}153$$

The predictions of this multiple linear regression model are represented by the gray line in Figure 10.1.

To assess the out-of-sample predictive accuracy of these two models, we compute the RMSE of their predictions on the four observations of the validation set (represented by the orange markers in Figure 10.1). Table 10.3 shows that based on this static validation set, the multiple linear regression model has a smaller RMSE.

If the validation set is used to select a final model through comparison with other models (as we did in Table 10.3), then the RMSE of this selected model tends to be an overly optimistic estimate of the model's accuracy in future predictions. Thus, to obtain an improved estimate of the final model's RMSE for future predictions we should apply the model to data that have not been used to construct or select the model. The data used to obtain unbiased estimates of the best model's future predictive performance are called the **test set**.

To demonstrate the use of a test set to obtain an unbiased estimate of a selected model's RMSE, consider the multiple regression model we identified as being the most accurate on the static validation set in Table 10.3, which uses the variables Age and Age^2. The first step is to aggregate the training set and the validation set, and the next step is to retrain the selected model on these 12 observations. We retrain the selected model on both the training and validation data because using this larger data set should result in the best-estimated model for predicting the future.

Figure 10.2 displays the estimated multiple linear regression model based on the 12 observations in the aggregated training and validation sets. Notice that the estimates of the slope coefficients ($-6{,}845$ for Age and 195 for Age^2) and the estimate of the constant (63,616) based on these 12 observations are slightly different from the estimates used for the multiple linear regression model fit to the eight observations in the static training set shown in Figure 10.1.

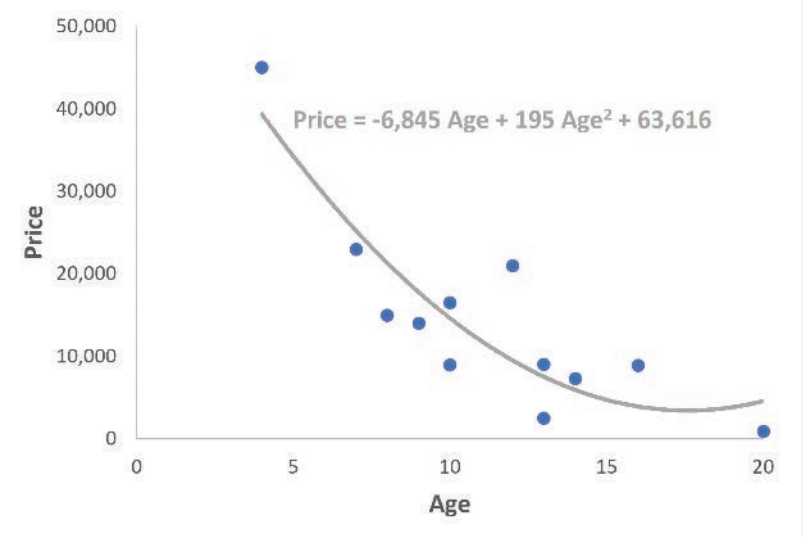
For a test set of four observations (set aside and not previously used to train or validate any regression model), Table 10.4 contains the predictions of the multiple linear regression model in Figure 10.2. For this test set, the RMSE of 18,576 provides a more conservative estimate of the multiple linear regression model's accuracy than the RMSE of 4,420 obtained on the static validation set. Because we selected the linear regression model using Age and Age^2 as our final model based on its RMSE of 4,420 on the validation set, this estimate will tend to be overly optimistic about our model's accuracy. Thus, applying the model to a test set provides a separate, unbiased estimate of the model's RMSE. Because the purpose of evaluating a model on a test set is to provide a final estimate of the model's future predictive performance, it needs to be sufficiently large to obtain a stable estimate.

Table 10.3 Calculation of RMSE on Static Validation Set

Car	Age (years)	Price ($)	Predicted Price from Simple Linear Regression	Error of Simple Linear Regression	Predicted Price from Multiple Linear Regression	Error of Multiple Linear Regression
9	10	8,950	18,139	−9,189	14,485	−5,535
10	16	8,888	3,986	4,902	2,662	6,226
11	10	16,500	18,139	−1,639	14,485	2,015
12	13	9,000	11,063	−2,063	6,832	2,168
			RMSE:	5,371	RMSE:	4,420

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 10.2 Selected Model from Static Holdout Retrained on Both Training and Validation Data



There are no definite rules for the size of the three partitions, but the training set is typically the largest. For a regression task, a rule of thumb for the size of the training set is to have at least 10 times as many observations as variables. In situations with relatively few observations, an analyst may not have the luxury of setting aside observations for a test set. In this case, the analyst may resort to using only training and validation sets. The lack of a test set does not impair the process of training and assessing candidate models as training and validation sets accomplish this. However, without applying a final model to a test set, we do not have an unbiased estimate of the final model's future performance and must rely on the optimistic estimate of RMSE from the validation set.

The advantages of the static holdout method are that it is simple and quick. However, the results of the modeling procedure based on a single application of the static holdout can vary greatly depending on which observations and how many observations are selected for each of the training and validation sets. If we would construct another iteration of the static holdout method and split the observations differently across the training and validation sets, we may find that a different model performs best. Motivated by this realization, we describe a more robust data partitioning method to train and validate models in the next section.

Table 10.4 Calculation of RMSE on Test Set

Car	Age (years)	Price ($)	Predicted Price from Multiple Linear Regression	Error of Multiple Linear Regression
13	6	23,495	29,552	−6,057
14	5	12,995	34,256	−21,261
15	7	32,999	25,237	7,762
16	3	15,999	44,832	−28,833
			RMSE:	18,576

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

k-Fold Cross-Validation

In the static holdout method, the observations that are placed into the training and validation sets can have an impact on how models are built and evaluated. In this section, we discuss a more robust procedure to train and validate models called ***k*-fold cross-validation.** As a pre-processing step, if an unbiased estimate of the final model is desired, we first remove a desired number of observations to compose the test set. The remaining observations will be used to train and validate the model in an iterative procedure. The *k*-fold cross-validation procedure randomly divides the data not in the test set into k approximately equal-sized, mutually exclusive, and collectively exhaustive subsets called folds. The cross-validation procedure consists of k iterations in which a different subset is designated as the validation set at each iteration and the remaining $k - 1$ subsets are combined and designated as the training set. At each iteration, a model is constructed using the respective training set data and evaluated using the respective validation set. At the end of the k iterations, the performance measures of the k models are aggregated and averaged. A common choice for the number of folds is $k = 5$ or 10. The *k*-fold cross-validation method is more complex and time consuming than the static holdout method, but the results of the *k*-fold cross-validation method are less sensitive to which observations are used in the training and the validation sets, respectively.

Leave-one-out cross-validation is a special case of *k*-fold cross validation in which each fold consists of a single observation. That is, for training and validation on a set of n observations, an iteration of leave-one-out cross-validation consists of estimating the model on $n - 1$ observations and evaluating the model on the single observation that was omitted from the training data. This procedure is repeated for n total iterations so that the model is trained on each possible combination of $n - 1$ observations and evaluated on the single remaining observation in each case.

To demonstrate the *k*-fold cross-validation procedure, let us again consider the 12 observations of pre-owned car age and sales price collected by Palanquin Auto. For these data, we will split these data into $k = 3$ folds with $12 \div 3 = 4$ observations in each fold. As shown in Table 10.5, fold 1 consists of cars 1–4, fold 2 consists of cars 5–8, and fold 3 consists of cars 9–12.

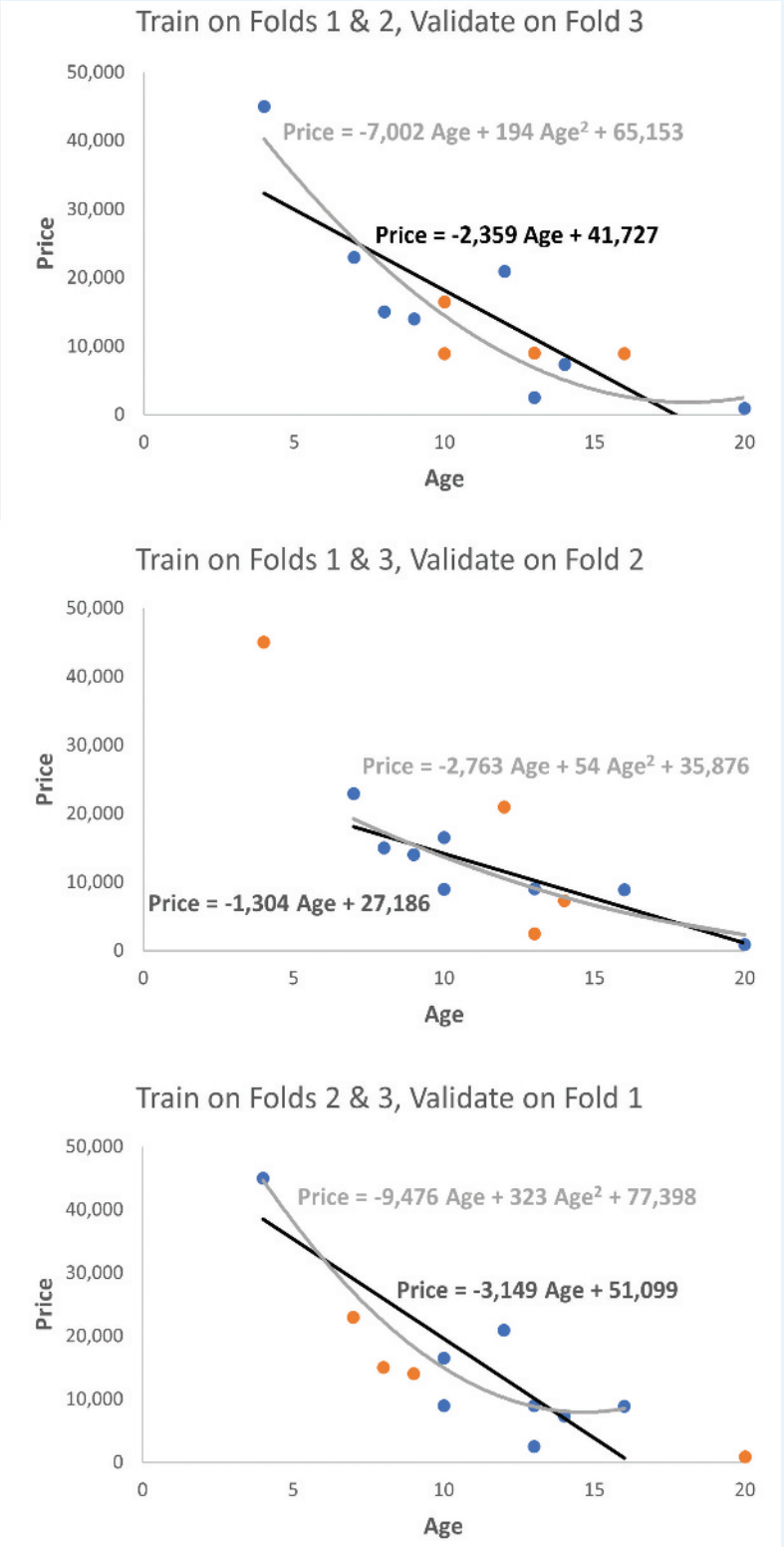
Using the three folds outlined in Table 10.5, Palanquin would like to train and assess two candidate regression models to predict the price of a pre-owned car. One model is a simple linear regression model using Age as the independent variable. The other model is a multiple linear regression model using Age and Age^2 as the independent variables. Figure 10.3

Table 10.5 Fold Cross-Validation for 12 Observations

Car	Age (years)	Price ($)	
1	9	14,000	Fold 1
2	8	15,000	
3	7	22,950	
4	20	900	
5	12	20,950	Fold 2
6	13	2,500	
7	14	7,300	
8	4	45,000	
9	10	8,950	Fold 3
10	16	8,888	
11	10	16,500	
12	13	9,000	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 10.3 Training (Blue) and Validation (Orange) Iterations in a 3-Fold Cross-Validation on Palanquin Data



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

illustrates the three iterations of training and validation corresponding to the 3-fold cross-validation procedure on the 12 observations from Palanquin Auto.

The top chart in Figure 10.3 displays the two models trained on the eight observations from folds 1 and 2 (represented by blue data points). Based on the eight observations of folds 1 and 2, an estimated simple linear regression model is

$$\text{Price} = -2{,}359\ \text{Age} + 41{,}727$$

The predictions of this simple linear regression model are represented by the black line in the top chart of Figure 10.3.

Based on the eight observations of folds 1 and 2, an estimated multiple linear regression model is

$$\text{Price} = -7{,}002\ \text{Age} + 194\ \text{Age}^2 + 65{,}153$$

The regression coefficients displayed in Figure 10.3 and the associated text are rounded, but the prediction and error calculations in Table 10.6 use unrounded regression coefficients.

The predictions of this multiple linear regression model are represented by the gray line in the top chart of Figure 10.3.

These two models are then used to predict the prices of the four cars in fold 3. For example, for car 9, the simple linear regression model predicts Price = −2,359 (10) + 41,727 = $18,139 and the multiple regression model predicts Price = −7,002 (10) + 194 (10^2) + 65,153 = $14,485. The predictions of the four cars in fold 3 using the two models trained on folds 1 and 2 are summarized in the top third of Table 10.6.

The middle chart in Figure 10.3 displays the two models trained on the eight observations from folds 1 and 3 (represented by blue data points). Based on the eight observations of folds 1 and 3, an estimated simple linear regression model is

$$\text{Price} = -1{,}304\ \text{Age} + 27{,}186$$

Table 10.6 Calculation of RMSE in a 3-Fold Cross-Validation

Car	Age (years)	Price ($)	Predicted Price from Simple Linear Regression	Error in Simple Linear Regression	Predicted Price from Multiple Linear Regression	Error in Multiple Linear Regression
Train on Folds 1 & 2, Validate on Fold 3						
9	10	8,950	18,139	−9,189	14,485	−5,535
10	16	8,888	3,986	4,902	2,662	6,226
11	10	16,500	18,139	−1,639	14,485	2,015
12	13	9,000	11,063	−2,063	6,832	2,168
Train on Folds 1 & 3, Validate on Fold 2						
5	12	20,950	11,534	9,416	10,533	10,417
6	13	2,500	10,230	−7,730	9,126	−6,626
7	14	7,300	8,926	−1,626	7,829	−529
8	4	45,000	21,969	23,031	25,692	19,308
Train on Folds 2 & 3, Validate on Fold 1						
1	9	14,000	22,758	−8,758	18,295	−4,295
2	8	15,000	25,907	−10,907	22,277	−7,277
3	7	22,950	29,056	−6,106	26,905	−3,955
4	20	900	−11,880	12,780	17,159	−16,259
			RMSE:	**9,972**	**RMSE:**	**8,921**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The predictions of this simple linear regression model are represented by the black line in the middle chart of Figure 10.3.

Based on the eight observations of folds 1 and 3, an estimated multiple linear regression model is

$$\text{Price} = -2{,}763\,\text{Age} + 54\,\text{Age}^2 + 35{,}876$$

The predictions of this multiple linear regression model are represented by the gray line in the middle chart of Figure 10.3.

These two models are then used to predict the prices of the four cars in fold 2. For example, for car 5, the simple linear regression model predicts Price = −1,304 (12) + 27,186 = \$11,534 and the multiple regression model predicts Price = −2,763 (12) + 54 (12^2) + 35,876 = \$10,533. The predictions of the four cars in fold 2 using the two models trained on folds 1 and 3 are summarized in the middle third of Table 10.6.

The bottom chart in Figure 10.3 displays the two models trained on the eight observations from folds 2 and 3 (represented by blue data points). Based on the eight observations of folds 2 and 3, an estimated simple linear regression model is

$$\text{Price} = -3{,}149\,\text{Age} + 51{,}099$$

The predictions of this simple linear regression model are represented by the black line in the bottom chart of Figure 10.3.

Based on the eight observations of folds 2 and 3, an estimated multiple linear regression model is

$$\text{Price} = -9{,}476\,\text{Age} + 323\,\text{Age}^2 + 77{,}398$$

The predictions of this multiple linear regression model are represented by the gray line in the bottom chart of Figure 10.3.

These two models are then used to predict the prices of the four cars in fold 1. For example, for car 1, the simple linear regression model predicts Price = −3,149 (9) + 51,099 = \$22,758 and the multiple regression model predicts Price = −9,476 (9) + 323 (9^2) + 77,398 = \$18,295. The predictions of the four cars in fold 1 using the two models trained on folds 2 and 3 are summarized in the bottom third of Table 10.6.

Because each of three iterations of the 3-fold cross-validation procedure considers a different collection of folds for its training set, the estimated regression coefficients for the simple linear regression model with Age as the independent variable and the multiple linear regression model with Age and Age^2 as independent variables vary from iteration to iteration. However, the functional form of these models is consistent regardless of which data are used to train the models. Thus, what we really desire to assess is how the structure of different models will perform on future observations.

In the remainder of this chapter, our examples demonstrating different predictive models for a regression task will use the static holdout method for simplicity. However, k-fold cross-validation is recommended in practice.

To assess the out-of-sample predictive accuracy of the simple linear regression model with Age as the independent variable and the multiple linear regression model with Age and Age^2 as independent variables, we consider the models' predictions on the different validation sets considered in the *k*-fold cross-validation procedure. In Figure 10.3, the four observations composing the validation set at each iteration of the cross-validation procedure are illustrated as orange data points. Due to the way the 3-fold cross-validation rotates folds between the training and validation sets, each of the 12 observations is part of the validation set exactly once. Thus, the models make out-of-sample predictions on each of 12 observations in the entire data exactly once. Table 10.6 displays these 12 predictions for both candidate models. In Table 10.6, the out-of-sample predictive accuracy is measured by computing the RMSE of the 12 predictions each model has made over the three iterations in the 3-fold cross-validation procedure. Based on the RMSE in this 3-fold cross-validation experiment, the multiple linear regression model is more accurate.

Because the models' RMSE on the validation sets of the 3-fold cross-validation procedure was used to select a final model, we should obtain an estimate of the final model's RMSE for future predictions by applying the model to a test set.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To demonstrate the use of a test set to obtain an unbiased estimate of a selected model's RMSE, consider the multiple linear regression model using Age and Age^2 as independent variables which we identified as being the most accurate in the 3-fold cross-validation in Table 10.6. The first step is to aggregate the training set and validation set and retrain the selected model on these 12 observations. At this point, we retrain the selected model on both the training and validation data because using this larger data set should result in the best estimated model for predicting the future. Figure 10.4 displays the estimated multiple linear regression model based on the 12 observations in the aggregated training and validation sets.

For a test set of four observations (set aside at beginning of analysis, and not previously used to train or validate any regression model), Table 10.7 contains the predictions of the multiple linear regression model in Figure 10.4. For this test set, the RMSE is 18,576 which provides a more conservative estimate of the multiple linear regression model's accuracy than the RMSE of 8,921 obtained from the 3-fold cross-validation. Because we selected the linear regression model using Age and Age^2 as our final model based on its RMSE 8,921 in the 3-fold cross-validation, this estimate will tend to be overly optimistic

Figure 10.4 Selected Model from 3-Fold Cross-Validation Retrained on Both Training and Validation Data

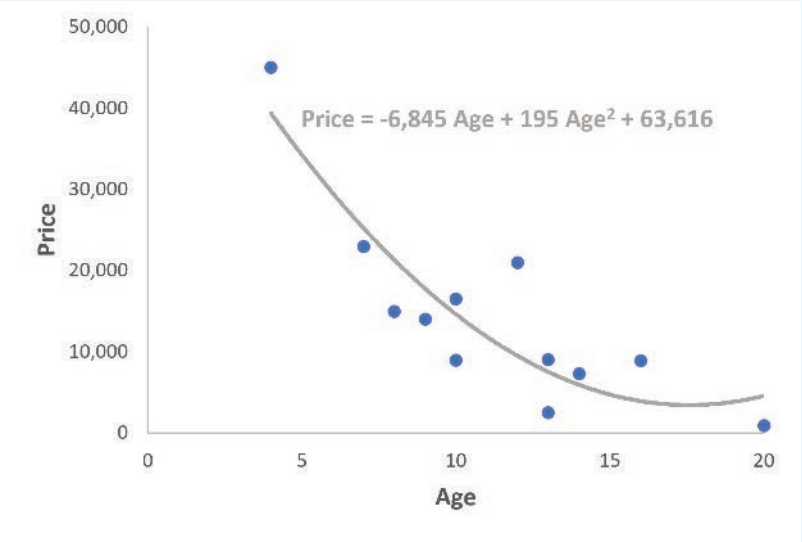


Table 10.7 Calculation of RMSE on Test Set

Car	Age (years)	Price ($)	Predicted Price from Multiple Linear Regression	Error
13	6	23,495	29,552	−6,057
14	5	12,995	34,256	−21,261
15	7	32,999	25,237	7,762
16	3	15,999	44,832	−28,833
			RMSE:	18,576

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

about our model's accuracy. Thus, applying the model to a test set provides a separate, unbiased estimate of the model's RMSE. Because the purpose of evaluating a model on a test set is to provide a final estimate of the model's future predictive performance, it needs to be sufficiently large to obtain a stable estimate.

Notes + Comments

The size of the training, validation, and test sets in our example are impractically small and are used here only as an illustrative example of the holdout process.

10.3 *k*-Nearest Neighbors Regression

For a regression task, ***k*-nearest neighbors (*k*-NN)** bases its prediction of an observation's quantitative outcome variable based on the k most similar observations from the training set, where similarity is measured with respect to the set of input variables (features). In the k-NN method, the predicted value of an observation's quantitative outcome variable is calculated to be the average of the values of the outcome variable for its k nearest neighbors in the training set.

The k-NN method is called a **lazy learner** because it performs no calculations on the training set prior to being supplied an observation to predict. The k-NN method does not learn model parameters from training data like linear regression, but instead identifies the k most similar observations from the training data and uses these k observations to make its prediction for a specified observation. There are two primary algorithm choices to specify for the k-NN method: how to measure the distance between observations and how many neighbors to consider in the calculation of the predicted outcome value. Data mining software commonly employs Euclidean distance in the k-NN method to measure the similarity between observations, but other distances measures can be used. The number of neighbors to consider (the value of k) can plausibly range from 1 to the number of observations in the training set. If $k = 1$, then the prediction of an observation's outcome value is set equal to the outcome value of the single most similar observation from the training set. At the other extreme, if k is equal to the number of observations in the training set, then the prediction of an observation's outcome value is equal to the average outcome value over the entire training set. Too small of a value for k results in predictions that are overfit to the noise in the training set, while too large of a value of k results in underfitting and fails to capture the relationships between the features and the outcome variable. The best value of k can be determined by building k-NN models over a range of k values and then selecting the value of $k^\star$ that results the smallest prediction error (as measured by RMSE or MAE) on validation data. Note that the use of validation data to identify $k^\star$ in this manner implies that the k-NN model should be applied to a test set with the value of $k^\star$ to accurately estimate the prediction error on future data.

We discuss Euclidean distance and other distance measures in more detail in Chapter 6.

DATA*file*
preowned12.csv

To illustrate k-NN regression, consider the regression task of predicting the sales price of a pre-owned car, using age measured in years (Age) and the miles on the odometer of the vehicle (Odom). Let the eight observations from Palanquin Auto in Table 10.8 compose the training set and designate the four observations in Table 10.9 as the validation set. For the Palanquin Auto example, k-NN regression will base its prediction of a pre-owned car's sales price based on the k most similar cars in the training set with respect to age and odometer mileage.

To prevent differences in the scale of the variables from creating disproportionate influence on the distance calculations in the k-nearest neighbors method, it is common practice to standardize the values of the variables. In this example, we standardize the values of Age and Odom for the observations by computing the respective z-scores. In the z-score calculations, we scale the variable values using the mean and standard deviation of

We discuss z-scores in Chapter 2.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the respective variable computed over the 8 observations in the training set. As Table 10.8 lists, the mean and standard deviation of the Age variable in the training set are 10.875 and 4.970, respectively. Thus, the standardized value of the Age variable for car 1 is

$$\text{car 1 Age (standardized)} = \frac{9 - 10.875}{4.970} = -0.3773$$

Similarly, the standardized value of the Odom variable for car 1 is

$$\text{car 1 Odom (standardized)} = \frac{95{,}000 - 150{,}237}{72{,}442} = -0.7625$$

Table 10.8 summarizes the standardized values of Age and Odom for the observations in the training set.

In a similar manner, we standardize the values of Age and Odom for the observations in the validation set. In the *z*-score calculations, we scale the variable values using the mean and standard deviation of the respective variable computed over the 8 observations in the training set. Note that we do not use the mean and standard deviation of the observations in the validation set in the *z*-score calculation of the observations in the validation set. By standardizing each variable in the validation set based on the mean and standard deviation of the respective variable in the training set, we avoid leaking any information about the validation set into the training of the *k*-NN regression model. Table 10.9 summarizes the standardized values of Age and Odom for the observations in the validation set.

Table 10.8 Training Set for *k*-NN Regression

Car	Age (years)	Odom (miles)	Age (standardized)	Odom (standardized)	Price ($)
1	9	95,000	–0.3773	–0.7625	14,000
2	8	50,500	–0.5785	–1.3768	15,000
3	7	166,380	–0.7797	0.2228	22,950
4	20	250,500	1.8362	1.3840	900
5	12	191,955	0.2264	0.5759	20,950
6	13	204,561	0.4276	0.7499	2,500
7	14	181,000	0.6288	0.4247	7,300
8	4	62,000	–1.3834	–1.2180	45,000
$\bar{x}$:	10.875	150,237	0	0	
s:	4.970	72,442	1	1	

Table 10.9 Validation Set for *k*-NN Regression

Car	Age (years)	Odom (miles)	Age (standardized)	Odom (standardized)	Price ($)
9	10	164,000	–0.1761	0.1900	8,950
10	16	53,320	1.0313	–1.3378	8,888
11	10	201,766	–0.1761	0.7113	16,500
12	13	56,700	0.4276	–1.2912	9,000

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In this example, we identify the *k*-nearest neighbors of an observation by computing the Euclidean distance between pairs of observations based on their respective (standardized) values of Age and Odom. The Euclidean distance between cars *u* and *v* is computed as

$$d_{uv}^{euclid} = \sqrt{(Age_u - Age_v)^2 + (Odom_u - Odom_v)^2}$$

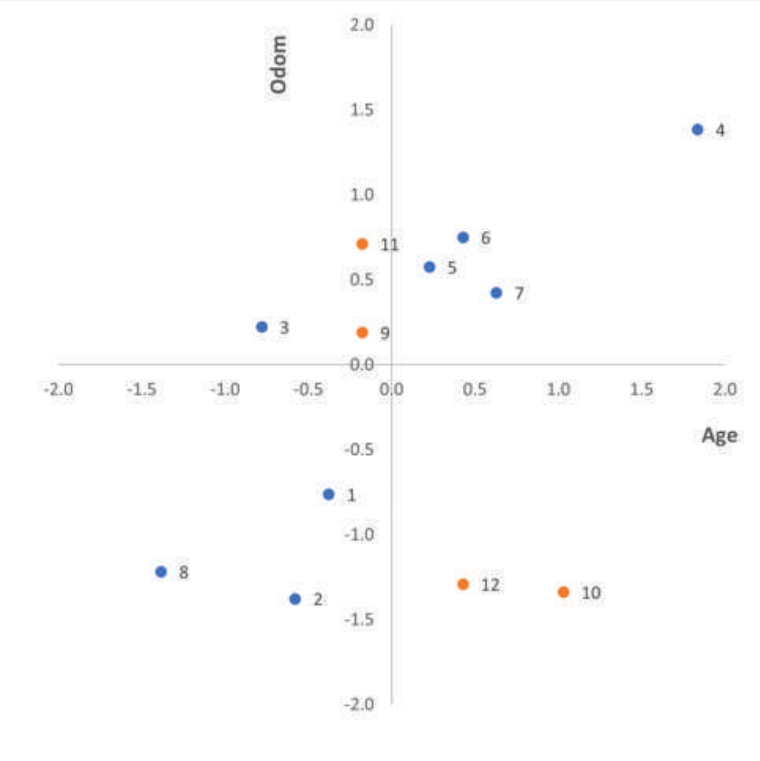
For example, consider car 1 in Table 10.8 and car 9 in Table 10.9. The Euclidean distance between these two cars is

$$d_{19}^{euclid} = \sqrt{(-0.3773 - (-0.1761))^2 + (-0.7625 - 0.1900)^2}$$
$$= 0.9735$$

Let's consider applying *k*-NN to predict the prices of the four cars in the validation set of Table 10.9 using $k = 3$. Thus, the price of a car in the validation set will be estimated by the average of the three cars in the training set that are the most similar to that car with respect to Age and Odom. To determine the $k = 3$ nearest neighbors for car 9 in the validation set, we compute the Euclidean distance between car 9 and each of the cars in the training set. These calculations are then repeated for the other cars in the validation set.

To visualize the *k* nearest neighbors in this example, Figure 10.5 plots the standardized training set observations as blue data points and the standardized validation set observations as orange data points. With respect to two variables, Euclidean distance between two observations corresponds to the length of the straight line these two observations. Using Figure 10.5, we can visually corroborate the $k = 3$ nearest training set neighbors for each of the observations in the validation set.

Figure 10.5 Standardized Training Set (Blue) and Validation Set (Orange) Observations



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 10.10 *k*-NN Predictions of Validation Set Observations

Car	Price ($)	$k = 3$ Closest Neighbors	Predicted Price ($)	Error
9	8,950	5, 3, 6	15,467	−6,517
10	8,888	1, 2, 7	12,100	−3,212
11	16,500	5, 6, 3	15,467	1,033
12	9,000	1, 2, 7	12,100	−3,100
			RMSE:	3,983

The three closest training set observations to car 9 are cars 5, 3, and 6. Thus, the *k*-NN prediction of car 9's price is $(20{,}950 + 22{,}950 + 2{,}500) \div 3 = 15{,}647$. The three closest training set observations to car 10 are cars 1, 2, and 7. Thus, the *k*-NN prediction of car 10's price is $(14{,}000 + 15{,}000 + 7{,}300) \div 3 = 12{,}100$. The *k*-NN prediction of the prices of cars 11 and 12 are calculated in a similar manner. Table 10.10 displays the three closest training set observations for all four validation set observations and the corresponding predicted price. The RMSE for the *k*-NN method with $k = 3$ on this validation set is 3,983. By varying *k* and repeating the predictions on validation data, the value of *k* resulting in the best predictive accuracy for the *k*-NN method can be determined.

Notes + Comments

1. In the standard *k*-NN regression method, the predicted value of an observation is computed as the simple average of the values of the outcome variable for the *k* nearest observations in the training set. In this version, each of the *k* nearest neighbors contributes uniformly to the predicted value. An option in many implementations of the *k*-NN algorithm is to compute the predicted value of an observation based on a weighted average of the *k* nearest observations in the training set, where the weights are inversely proportional to the distance to the observation being predicted. In this weighted version, the value of the outcome variable for a neighboring observation that is closer to the observation being predicted is given more weight in the predicted value.

10.4 Regression Trees

A **regression tree** is a predictive model consisting of a series of rules on the input features so that each successive rule splits the training set of observations into smaller subsets that are more homogeneous with respect to the quantitative outcome variable. In this section, we begin by describing how a regression tree is constructed based on a training set. Then we discuss using the regression tree to make predictions and pruning a regression tree to prevent overfitting.

Constructing a Regression Tree

Because there are several different algorithms for constructing a regression tree, various data mining software packages may obtain different trees. In this section, we present a general description of tree-based logic which may slightly differ from the implementation of regression trees in a specific data mining software.

At each iteration in the construction of a regression tree, a subset of observations is further divided into two new subsets using a threshold rule based on the observations' values of a single input variable. The input variable selected for the rule and the threshold value of this input variable upon which to divide the observations is determined to decrease the aggregate variance of the quantitative outcome variable in the two resulting subsets as much as possible. Thus, the sequence of rules in a regression tree can be thought of a series of questions that successively split observations into smaller groups so that the aggregate variance of the quantitative outcome variable in the resulting groups is as small as possible.

To demonstrate the construction of a regression tree, we consider the task of predicting pre-owned car prices for Palaquin Auto. Table 10.11 contains the eight observations we consider as the training set for the regression tree. In our explanation, we will refer to Figures 10.6 and 10.7. Figure 10.6 displays the full regression tree obtained when applying a

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
preowned12.csv

Table 10.11 Training Set of Palaquin Data for Regression Tree

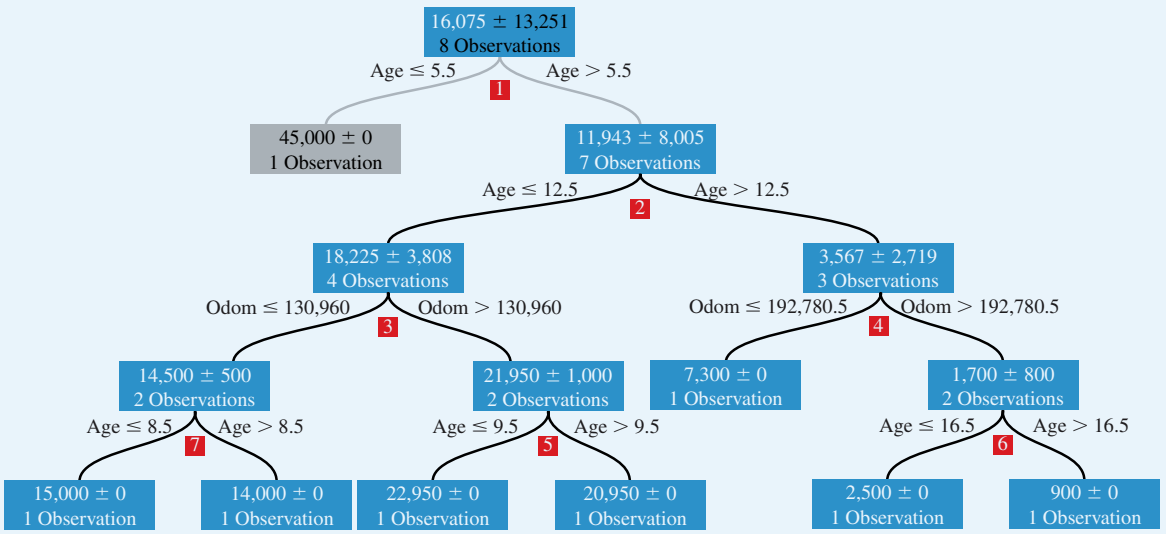
Car	Age (years)	Odom (miles)	Price ($)
1	9	95,000	14,000
2	8	50,500	15,000
3	7	166,380	22,950
4	20	250,500	900
5	12	191,955	20,950
6	13	204,561	2,500
7	14	181,000	7,300
8	4	62,000	45,000

In the Palanquin example, the input variables are quantitative. When considering a categorical input variable, rather than considering possible threshold values, the splitting procedure will consider the possible ways to split the categories of the input variable into two groups.

set of rules that segments the set of observations to the greatest extent possible. Figure 10.7 contains a scatter chart that provides a geometric illustration of the full regression tree in Figure 10.6. In Figure 10.7, the coordinates of each blue data point correspond to the values of Age and Odom for a training set observation. Each of the blue data points is labeled with the observation's value of Price.

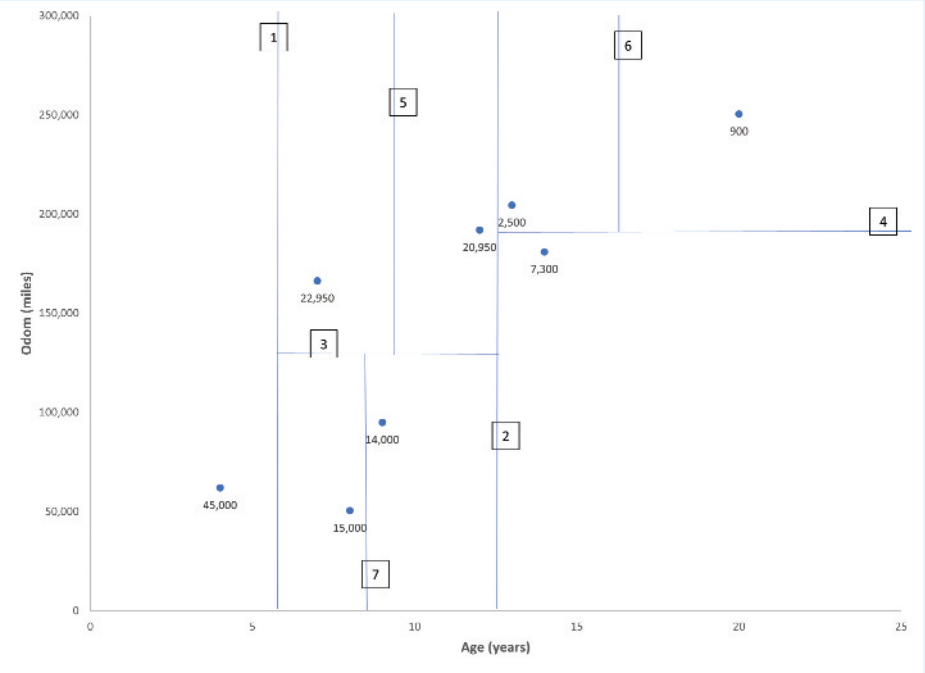
Each rule in a regression tree splits a group of observations into two subsets based on the value of a single input variable. As Figure 10.6 shows, the subsets of observations are displayed as nodes in a regression tree. The top node in the regression tree of Figure 10.6 is called the **root node** (a regression is an upside-down tree with the roots at the top). The root node in Figure 10.6 conveys that in the entire training set there are eight observations with an average price of $16,075 and price standard deviation of $13,251. We then consider all possible ways to split these eight observations into two separate groups based on their values of Age or Odom, in order to obtain the largest decrease in the aggregate variance in Price. The largest reduction in variance results from splitting the eight observations into two groups: a group with a single observation with Age $\leq$ 5.5 and another group with seven observations with Age $>$ 5.5. The left branch emanating from the root node in Figure 10.6 corresponds to the group with the single observation with Age $\leq$ 5.5. The right branch emanating from the root node in Figure 10.6 corresponds to the group with seven observations with Age $>$ 5.5.

Figure 10.6 Full Regression Tree on Palanquin Data



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 10.7 Geometric Illustration of Full Regression Tree Segmentation of the Input Variable Space



Before the first split, the variance of Price for the eight observations is $13{,}251^2 = 175{,}588{,}750$. After splitting these eight observations into a group with a single observation with a price standard deviation of 0 and a group of seven observations with a price standard deviation of 8,005, the aggregate variance of Price for these two groups is

$$\frac{(1 \times 0^2) + (7 \times 8{,}005^2)}{8} = 56{,}066{,}518$$

Thus, by splitting the eight observations on the rule considering variable Age and the value 5.5, we reduce the price variance by $(175{,}588{,}750 - 56{,}066{,}518) = 119{,}522{,}232$. Viewing the training set of observations on a scatter chart of the input variables Age and Odom, we can geometrically visualize that the rule based on the variable Age and the value 5.5 as the vertical line in Figure 10.7 labeled with 1.

The left branch emanating from the root node in Figure 10.6 corresponding to the group with Age ≤ 5.5 only contains a single observation with a price of \$45,000 and a price standard deviation of 0, so no further rules to further split this group are required. However, the right branch emanating from the root node in Figure 10.6 corresponding to the seven observations with Age > 5.5 can be split further. This group of seven observations has an average price of \$11,943 and a price standard deviation of \$8,005.

In step 2, the regression tree considers all possible ways to further split the seven observations with Age > 5.5 into two separate groups based on their values of Age or Odom in order to obtain the largest decrease in the aggregate within-group variance of Price. As Figure 10.6 shows, the seven observations with Age > 5.5 are split into a group of four observations with Age ≤ 12.5 and a group of three observations with Age > 12.5. Thus, splitting the seven observations with Age > 5.5 into two groups using the variable Age and the value 12.5, we obtain the largest reduction in price variance of these seven observations. Geometrically, the rule based on the variable Age and the value 12.5 corresponds to the vertical line in Figure 10.7 labeled with 2.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

After these first two steps of the regression tree, the original eight observations have been split into three total groups: one observation with Age $\leq$ 5.5, four observations with Age $>$ 5.5 and Age $\leq$ 12.5, and three observations with Age $>$ 12.5.

Following a similar process, the group of four observations with Age $>$ 5.5 and Age $\leq$ 12.5 is further split into two groups: a group of two observations with Odom $\leq$ 130,960 and a group of two observations with Odom $>$ 130,960. Geometrically, the rule based on the variable Odom and the value 130,960 corresponds to the horizontal line in Figure 10.7 labeled with 3.

The process of generating rules to further split groups of observations continues in the order denoted by the boxed numbers in Figure 10.6. The process stops at a **leaf node**, a node in a tree that is not further split, when the variance of a group of observations can no longer be reduced by a rule on a value of one of the input variables. In this example, this occurs when a group of observations has zero variance in Price. In general, non-zero variance may occur in a leaf node when multiple observations have the same set of values for the input variables, but different values of the outcome variable. For this example, that would correspond to a case in which two cars have the same age and odometer reading, but different sales prices.

Figure 10.7 illustrates that when considering only two input variables, a regression tree segments the observations into rectangular regions. The full regression tree corresponding to Figure 10.7 segments the observations until the variance in Price in each rectangular region is small as possible.

Generating Predictions with a Regression Tree

After a final regression tree is constructed from a training set, it can be used to predict observations in a validation set. For each observation in a validation set, the tree's rules are applied to the observation's input features to determine which subset of training set observations that the validation set observation belongs. Then the average outcome value of this subset of the training set observations is the predicted value of the outcome variable for the observation in the validation set.

Returning to the Palanquin example, let us first consider the prediction accuracy of the full regression tree in Figure 10.6. Recall that the full regression tree was constructed to partition to the fullest extent the eight observations of the training data. Let us now assess how well this full regression tree predicts the four validation set observations in Table 10.12.

Table 10.12 Validation Set Prediction Results for Full Regression Tree

Car	Age (years)	Odom (miles)	Price ($)	Branch Sequence	Predicted Price ($)	Error
9	10	164,000	8,950	Age $>$ 5.5, Age $\leq$ 12.5, Odom $>$ 130,960, Age $>$ 9.5	20,950	−12,000
10	16	53,320	8,888	Age $>$ 5.5, Age $>$ 12.5, Odom $<$ 192,780.5	7,300	1,588
11	10	201,766	16,500	Age $>$ 5.5, Age $\leq$ 12.5, Odom $>$ 130,960, Age $>$ 9.5	20,950	−4,450
12	13	56,700	9,000	Age $>$ 5.5, Age $>$ 12.5, Odom $<$ 192,780.5	7,300	1,700
					RMSE:	6,504

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To generate a prediction from a regression tree for a validation set observation, we must determine which leaf node contains the observation. Consider car 9 from the validation set in Table 10.12 with Age = 10 and Odom = 164,000. Starting at the root node labeled with a 1 in Figure 10.6, we travel down the right branch because Age = 10 > 5.5. At the next node labeled with a 2, because Age = 10 ≤ 12.5 we travel down the left branch. At the next node labeled with a 3, because Odom = 164,000 > 130,960 we travel down the right branch. At the next node labeled with a 5, because Age = 10 > 9.5 we travel down the right branch to land at the leaf node with the predicted value of $20,950. This results in a prediction error of 8,950 − 20,950 = −12,000. Table 10.12 summarizes the branch sequence and predicted price for all four validation set observations. The RMSE of the predictions from the full regression tree is $6,504.

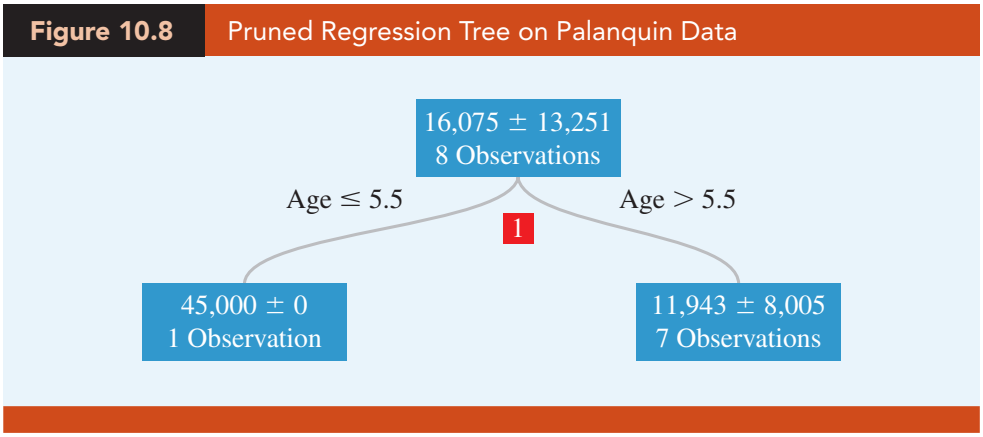
With its extensive segmentation of the training set observations, a full regression tree explains the training data as well as possible, but may not result in the most accurate out-of-sample predictions if it is overfit to the training data. This overfitting may occur because the predicted values generated by the leaf nodes are based on very few observations. For example, the eight leaf nodes in Figure 10.6 each only contain one training set observation. Thus, the predicted price of each validation set observation in Table 10.12 is based on only one training set observation. This suggests that a regression tree with fewer branches may produce more accurate out-of-sample predictions.

The process of obtaining a regression tree with fewer branches than a full regression tree is called **pruning**. There are several ways to prune a full regression tree to obtain a final regression tree with the best possible out-of-sample predictive accuracy. One approach is to sequentially remove branches from the bottom of the tree and assess the out-of-sample predictive performance of the sequence of increasingly smaller regression trees. Starting from the full regression tree in Figure 10.6 with eight leaf nodes, this would result in testing the predictive accuracy of a regression tree with seven leaf nodes, then six leaf nodes,…, and finally only one leaf node (the root node).

Another approach to reducing the size of a regression tree is to specify restrictions on how much a rule can split up a group of observations. One common restriction is to specify that each of the two subsets resulting from a variable splitting rule must have at least a specified number of observations. Yet another common restriction is to prohibit the splitting of a group of observations unless it has more than a specified number of observations.

Regardless of the approach used to prune a regression tree, the appropriate size of the tree is determined through assessing its out-of-sample predictive accuracy. For example, the pruned regression tree that obtains the lowest RMSE on the four validation set observations is displayed in Figure 10.8. As shown in Table 10.13, this pruned regression tree achieves a RMSE of $3,454 on the four validation set observations compared to the RMSE of $6,504 for the full regression tree.

Figure 10.8 Pruned Regression Tree on Palanquin Data



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 10.13 Validation Set Prediction Results for Pruned Regression Tree

Car	Age (years)	Odom (miles)	Price ($)	Branch Sequence	Predicted Price ($)	Error
9	10	164,000	8,950	Age > 5.5	11,943	−2,993
10	16	53,320	8,888	Age > 5.5	11,943	−3.055
11	10	201,766	16,500	Age > 5.5	11,943	4,557
12	13	56,700	9,000	Age > 5.5	11,943	−2,943
					RMSE:	3,454

Ensemble Methods

In the previous section, we demonstrated the prediction of an observation's value of a quantitative outcome variable based on the rules of a single regression tree. In this section, we discuss making a prediction for a regression task by combining the predictions of several regression trees. Combining a collection of regression trees to make a composite prediction is an example of an ensemble method. An **ensemble method** is a predictive method that combines the predictions from a collection of individual predictive models to generate its prediction of the value for an observation's outcome variable.

To gain an intuitive grasp of why an ensemble of prediction models may outperform, on average, any single prediction model, let's consider the task of predicting the value of the S&P 500 Index one year in the future. Suppose there are 100 financial analysts independently developing their own forecast based on a variety of information. One year from now, there certainly will be at least one analyst whose forecast will prove to be the most accurate. However, identifying beforehand which of the 100 analysts will be the most accurate may be virtually impossible. Therefore, instead of trying to pick one of the analysts and depending solely on their forecast, an ensemble approach would combine their forecasts by taking an average of the 100 forecast values and use this as the predicted value of the S&P 500 Index. The two conditions for an ensemble to perform better than a single model are as follows: (1) The individual models are constructed independently of each other, and (2) the individual models perform better than just randomly guessing.

This example also describes a notion known as "wisdom of the crowds," which contends that the aggregated knowledge of a group of people is generally superior to the knowledge of any one individual.

There are two primary steps to an ensemble approach: (1) the construction of a committee of individual models, and (2) the combination of the individual models' predictions to form a composite prediction. An ensemble approach tends to be most effective when the individual models under consideration are unstable. A predictive model is an **unstable learner** if relatively small changes in the training set cause its predictions to fluctuate substantially. In this section, we discuss ensemble methods using regression trees, which are known to be unstable. Specifically, we discuss three different ways to construct an ensemble of regression trees: bagging, boosting, and random forests.

In the **bagging** approach, the committee of individual models is generated by first constructing multiple training sets by repeated random sampling of the n observations in the original data *with replacement*. Because the sampling is done with replacement, this process may result in training sets which contain multiple copies of some observations and no copies of other observations. The bagging approach then trains a regression model on each of the m training sets and generates the ensemble prediction based on the average of the m individual predictions. A basic premise of bagging is that training unstable learners on the slightly different m training sets will generate diversity in the resulting predictions.

DATA*file*
preowned12.csv

To demonstrate bagging, we consider the task of predicting the sales price of a pre-owned car only using information about the age of the car. Table 10.14 contains eight training set observations from the Palanquin Auto example. In Table 10.14, each of the eight columns

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 10.14 Original Training Set of 8 Observations

Car	8	3	2	1	5	6	7	4
Age (years)	4	7	8	9	12	13	14	20
Price ($)	45,000	22,950	15,000	14,000	20,950	2,500	7,300	900

corresponds to a pre-owned car observation and these observations (columns) have been sorted in increasing order of age to facilitate the following discussion of bagging. Table 10.15 shows the results of generating 10 new training sets by randomly sampling from the observations in Table 10.14 with replacement. For each of these training sets, we construct a one-rule regression tree (called a **stump**) that minimizes the aggregate variance of Price in the resulting two groups. The two groups of each training sample are denoted with a vertical red line and accompanying decision rule.

Table 10.16 shows the results of applying this ensemble of 10 regression trees to a validation set consisting of 4 observations. The ensemble method bases its prediction on the average of the 10 individual regression trees' predictions. In Table 10.16, the RMSE of the individual regression trees ranges from 3,411 to 10,384 while the RMSE of the ensemble is 2,880. Thus, on these four validation set observations, the ensemble is more accurate than any of the individual regression trees in the collection.

As an alternative to using a separate validation set (as the one in Table 10.16), the predictive performance of the bagging ensemble constructed in Table 10.15 can be assessed by using **out-of-bag estimation**. Out-of-bag estimation leverages the fact that because the training sets for each individual model used in the ensemble is trained by sampling the data with replacement, each of these training sets will each be missing an average of about 36.8% of the distinct observations.[1] Each of these missing observations can be used as holdout data upon which to evaluate the subset of models which do not contain the missing observations.

To demonstrate out-of-bag estimation, consider the bagging ensemble in Table 10.15. Because the training sets for each respective tree in Table 10.15 are constructed by sampling the original 10 observations in Table 10.14 with replacement, these training sets do not contain all 10 original observations. For example, the training set for Tree 1 does not contain the four observations:

Car	Age	Price
5	12	20,950
6	13	2,500
7	14	7,300
4	20	900

Therefore, these four observations constitute holdout data for Tree 1. For each tree in the ensemble method, Table 10.17 shows the price predicted by the respective tree for observations that were missing from the respective tree's training set. For example, Tree 1's rule is to predict a price of $45,000 for all cars with Age $\leq$ 5.5 and a price of $16,700 for all cars with Age $>$ 5.5. Therefore, Tree 1 predicts a price of $16,700 for each of the four observations with Age $=$ 12, 13, 14, and 20, respectively.

After each tree predicts the observations missing from its respective training set, the out-of-bag prediction of each of the original 10 observations is executed by averaging the predictions of the trees participating on each respective observation. For example, the observation with Age $=$ 4 is missing from the training sets of Trees 5 and 8. For the observation with Age $=$ 4, Tree 5 predicts a price of $16,328 and Tree 8 predicts a price of $15,421. Thus, the out-of-bag

[1]For a data set with n observations, the probability of an observation not being selected when randomly sampling the data set with replacement is $\left(\frac{n-1}{n}\right)^n$. For large n, $\left(\frac{n-1}{n}\right)^n$ is approximately 0.368.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 10.15 Bagging: Generation of 10 New Training Samples and Corresponding Regression Trees

Sample 1/Tree 1					Age ≤ 5.5			
Age	4	7	7	8	9	9	9	9
Price	45,000	22,950	22,950	15,000	14,000	14,000	14,000	14,000
Prediction	45,000				16,700			
Sample 2/Tree 2					Age ≤ 5.5			
Age	4	4	7	7	12	13	14	14
Price	45,000	45,000	22,950	22,950	20,950	2,500	7,300	7,300
Prediction	45,000				13,992			
Sample 3/Tree 3					Age ≤ 5.5			
Age	4	7	9	9	12	13	14	20
Price	45,000	22,950	14,000	14,000	20,950	2,500	7,300	900
Prediction	45,000				11,800			
Sample 4/Tree 4					Age ≤ 6.0			
Age	4	8	9	12	20	20	20	20
Price	45,000	15,000	14,000	20,950	900	900	900	900
Prediction	45,000				7,650			
Sample 5/Tree 5					Age ≤ 11			
Age	7	9	9	9	13	14	14	20
Price	22,950	14,000	14,000	14,000	2,500	7,300	7,300	900
Prediction			16,238				4,500	
Sample 6/Tree 6					Age ≤ 5.5			
Age	4	7	8	9	14	14	20	20
Price	45,000	22,950	15,000	14,000	7,300	7,300	900	900
Prediction	45,000				9,764			
Sample 7/Tree 7					Age ≤ 5.5			
Age	4	4	7	8	8	12	13	20
Price	45,000	45,000	22,950	15,000	15,000	20,950	2,500	900
Prediction	45,000				12,883			
Sample 8/Tree 8					Age ≤ 13			
Age	8	8	8	9	9	9	12	14
Price	15,000	15,000	15,000	14,000	14,000	14,000	20,950	7,300
Prediction				15,421				7,300
Sample 9/Tree 9					Age ≤ 12.5			
Age	4	7	7	8	12	13	13	13
Price	45,000	22,950	22,950	15,000	20,950	2,500	2,500	2,500
Prediction			25,370				2,500	
Sample 10/Tree 10					Age ≤ 6.5			
Age	4	9	12	14	14	14	20	20
Price	45,000	14,000	20,950	7,300	7,300	7,300	900	900
Prediction	45,000				8,379			

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 10.16 Prediction of 10 Observations from Validation Set with Bagging Ensemble

Car	9	10	11	12	
Age (years)	10	16	10	13	
Price ($)	8,950	8,888	16,500	9,000	RMSE
Tree 1	16,700	16,700	16,700	16,700	6,716
Tree 2	13,992	13,992	13,992	13,992	4,546
Tree 3	11,800	11,800	11,800	11,800	3,411
Tree 4	7,650	7,650	7,650	7,650	4,565
Tree 5	16,238	4,500	16,238	4,500	4,814
Tree 6	9,764	9,764	9,764	9,764	3,442
Tree 7	12,883	12,883	12,883	12,883	3,860
Tree 8	15,421	7,300	15,421	15,421	4,658
Tree 9	25,370	2,500	25,370	2,500	10,384
Tree 10	8,379	8,379	8,379	8,379	4,091
Ensemble Average:	13,820	9,547	13,820	10,359	2,880

prediction for the observation with Age = 4 is (16,328 + 15,421)/2 = $15,829. The error of this prediction is 45,000 − 15,829 = 29,171. The out-of-bag predictions for the other seven observations in the original training data are generated similarly. From Table 10.17, we see that the out-of-bag estimate of RMSE is 12,802. While this out-of-bag estimate of the RMSE is much higher than the RMSE of 2,880 observed on the validation set in Table 10.16, these values are typically much closer for larger ensembles. However, it is important to note that out-of-bag estimation does not assess performance as thoroughly as a separate validation set or a cross-validation procedure. In general, out-of-bag estimates will be more conservative than the analogous

Table 10.17 Out-of-Bag Predictions of Observations Missing from Each Tree's Training Set

Age (years)	4	7	8	9	12	13	14	20	
Price ($)	45,000	22,950	15,000	14,000	20,950	2,500	7,300	900	
Tree 1					16,700	16,700	16,700	16,700	
Tree 2			13,992	13,992				13,992	
Tree 3			11,800						
Tree 4		7,650				7,650	7,650		
Tree 5	16,238		16,238		4,500				
Tree 6					9,764	9,764			
Tree 7				12,883			12,883		
Tree 8	15,421	15,421				15,421		7,300	
Tree 9				25,370			2,500	2,500	
Tree 10		8,379	8,379			8,379			
OOB Average:	15,829	10,483	12,602	17,415	10,321	11,583	9,933	10,123	RMSE:
OOB Error:	29,171	12,467	2,398	−3,415	10,629	−9,083	−2,633	−9,223	12,802

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

performance measures based on a validation set. That is, an out-of-bag error estimate will be larger than a validation-based error estimate. Looking down the columns of Table 10.17, we observe that the out-of-bag prediction of each observation is based on a different subset of the trees.[2] Looking across the rows of Table 10.17, we observe that the trees participate in varying numbers of out-of-bag predictions.[3] In a separate validation set, each observation is predicted by each tree in the ensemble. As illustrated in Table 10.17, between two to five trees participate in the out-of-bag prediction of the eight original observations in the training data. However, in Table 10.16, all 10 trees contribute a prediction to the ensemble's composite predictions on four observations in the separate validation set.

Similar to bagging, the **boosting** method generates its committee of individual models by sampling multiple training sets. However, boosting differs from bagging in how it samples the multiple training sets and how it weights the individual models' predictions to generate the ensemble's composite prediction. Unlike bagging, in which the generation of the training sets (and corresponding individual models) are independent of each other, in boosting the construction of the *jth* training set (and the corresponding model) is dependent on the previous $j - 1$ training sets and models. Boosting iteratively adapts the construction of the *jth* training set based on the prediction error of the models constructed on the previous $j - 1$ training sets. There are several different implementations of boosting, but we describe one common approach here. In this implementation, to generate the first training set, each of the n observations in the original data is initially given equal weight of being selected. That is, each observation i has weight $w_i = 1/n$. A regression tree model is then constructed on this training set and is used to predict the outcome of the n observations in the original data. The weight of each observation i is then adjusted based on the degree of its prediction error. The next training set is then generated by sampling the observations according to the updated weights. In this manner, the next training set is more likely to contain observations that have been mispredicted in early iterations.

To combine the predictions of the m individual models from the m training sets, boosting weights the vote of each individual model based on its overall prediction error. For example, suppose that the classifier associated with the jth training set has a large prediction error and the classifier associated with the kth training set has a small prediction error. Then the classification votes of the jth classifier will be weighted less than the classification votes of the kth classifier when they are combined. Note that this method differs from bagging, in which each of the individual classifiers has an equally weighted vote.

A **random forest** is an extension of bagging specifically tailored for use with regression trees. As in bagging, the random forests approach generates multiple training sets by randomly sampling (with replacement) the n observations in the original data. However, when constructing a tree model for each separate training set, each tree is restricted to using only a specified number of randomly selected input variables. The individual regression trees are referred to as "weak learners" because they are only allowed to consider a small subset of input variables. A **weak learner** is a predictive model whose individual predictions are only slightly better than random guessing. Ensembles commonly consist of unstable learners that are intentionally weakened in an attempt to improve the independence of their predictions. Pruning the individual trees in a random forest is not typically necessary due to their restricted choice of input variables and their incorporation into an ensemble reduces the likelihood of overfitting. While the best number of individual trees in the random forest depends on the data, it is not unusual for a random forest to consist of hundreds and even thousands of individual trees.

For most data, the predictive performance of boosting ensembles exceeds the predictive performance of bagging ensembles. Boosting achieves its performance advantage because: (1) it evolves its committee of models by focusing on observations that are mispredicted, and

[2]It is possible that an observation is included in every base model's training set. In this case, the observation would not contribute to the out-of-bag estimation process.

[3]It is possible that a tree's training set contains every distinct observation and therefore has no original observations missing from its training set. In this case (although unlikely), the tree would not contribute to the out-of-bag estimation process.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(2) individual models' predictions are weighted by their accuracy. However, boosting is more computationally expensive than bagging. Because there is no adaptive feedback in a bagging approach, all m training sets and corresponding models can be implemented simultaneously. However, in boosting, the first training set and predictive model guide the construction of the second training set and predictive model, and so on. The random forests approach has performance similar to boosting, but maintains the computational simplicity of bagging.

10.5 Neural Network Regression

A **neural network** is a predictive data mining method whose structure is motivated by the biological functioning of the brain. In this section, we describe how a neural network makes predictions about a quantitative outcome variable, y, by processing values of a set of input variables, $x_1, x_2, \ldots, x_q$, through a series of calculations involving estimated parameters and the application of functions. We begin by describing the structure of a neural network and how this structure is used to generate a prediction. Then we describe how a neural network estimates its parameter values from data to model complex relationships between the input features and the outcome variable.

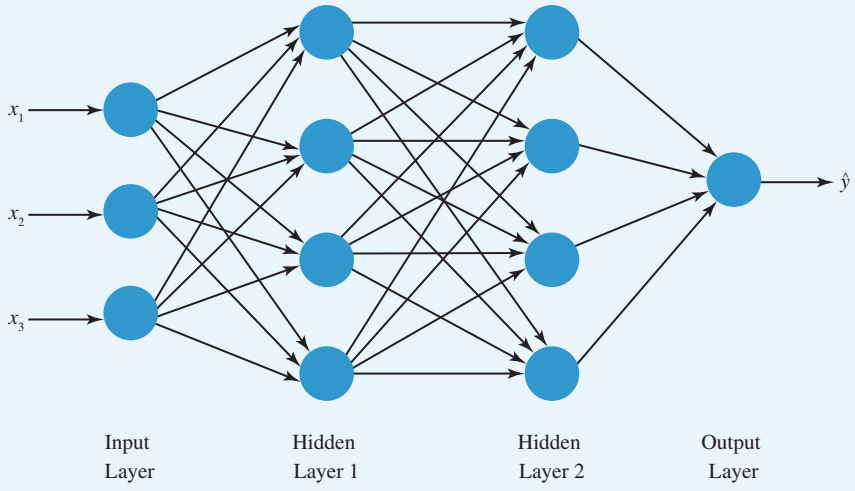
Structure of a Neural Network

A neural network consists of a layered system of nodes called **neurons**. Figure 10.9 illustrates a neural network with four layers: an input layer, two hidden layers, and an output layer. The **input layer** is set of neurons in which values of the input features enter the neural network. In Figure 10.9, the input layer consists of three neurons, corresponding to the input features x_1, x_2 and x_3. Typically, the number of neurons in the input layer corresponds to the number of input features.

Each layer between the first layer (the input layer) and the last layer (the output layer) is called a **hidden layer**. In the construction of a neural network, the number of hidden layers and the number of neurons in each hidden layer is an important design decision. As the number of hidden layers and number of neurons in each hidden layer increases, the neural network gains the capability of capturing more complex relationships between the input features and the outcome variable. However, the amount of data required to estimate the parameters reflecting these relationships also increases.

The last layer in a neural network is called the **output layer**. The number of neurons in the output layer corresponds to the number of different outcome variables being predicted. For regression tasks we consider in this chapter, the output layer will consist of a single neuron.

Figure 10.9 A Feed Forward Neural Network



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The neural network in Figure 10.9 is called a **feed forward neural network** because the prediction calculations are processed through connections between neurons in only one direction—from the input layer in which the values of the input variables enter, through the hidden layer(s) of the network, until the predicted value of the outcome variable is generated at the output layer. In this section, we only consider feed forward neural networks.

We begin our discussion of how a neural network generates its predictions by considering the simplest form of neural network. A **perceptron** is a neural network with no hidden layers (only an input layer and output layer). Figure 10.10 contains a perceptron with two input neurons connected to a single output neuron. Figure 10.10 displays the basic components of a neural network: neuron weights (w_{13},w_{23}), a bias term (b_3), and an activation function (f). In Figure 10.10, w_{13} is the weight applied to the neuron corresponding to feature x_1 and w_{23} is the weight applied to the neuron corresponding to feature x_2. A **neuron weight** represents the strength of a neuron's output signal to a neuron in the next layer. In Figure 10.10, a bias term b_3 is associated with the output neuron. The **bias term** amplifies (if $b_3 > 0$) or dampens (if $b_3 < 0$) the input signal into a neuron. At the output neuron, the predicted value of the outcome variable is obtained by computing $f(w_{13}x_1 + w_{23}x_2 + b_3)$ for a specified activation function f. An **activation function** is a function that produces an output from a neuron based on inputs into a neuron.

DATA *file*
preowned12.csv

Let's consider the application of the perceptron to the Palanquin Auto example. Analysts at Palanquin are interested predicting the sales price of a pre-owned car using information on the vehicle's age (Age) and odometer reading (Odom). To compute a prediction using the perceptron in Figure 10.10, we need to specify values of the neuron weights (w_{13},w_{23}), the bias term (b_3), and the form of the activation function (f). In the next subsection, we describe how a neural network learns values of the neuron weights and bias terms to improve the accuracy of its predictions, but for now, suppose we know these values. Figure 10.11 displays a version of the perception in which $w_{13} = -0.2457$,$w_{23} = -0.2728$, $b_3 = 0.3611$, and the activation function is the identity function, $f(x) = x$ which here means that we sum the

Figure 10.10 Perceptron Model

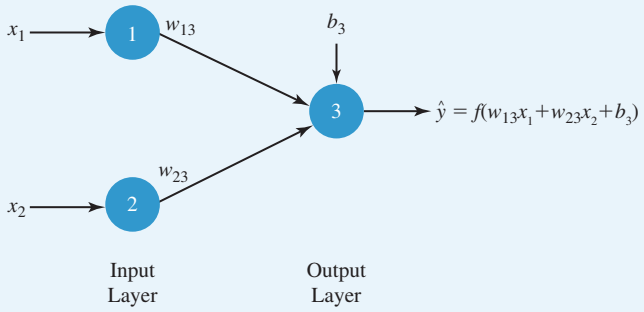
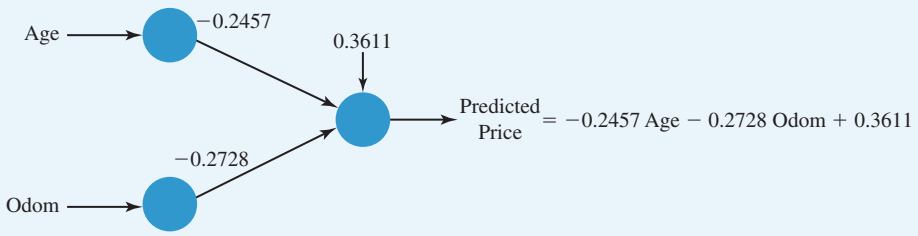


Figure 10.11 Perceptron Model for Palanquin Auto Using Identity Activation Function



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

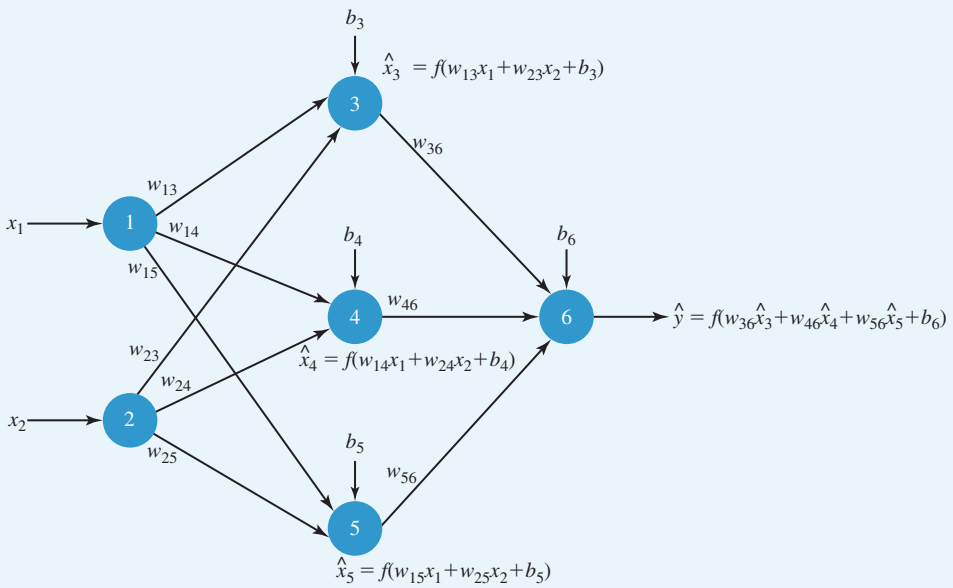
weighted inputs and the bias term. As Figure 10.11 shows, a perceptron model implemented using the identity function as the activation function results in a linear prediction equation for price. So, in its simplest form, a neural network regression has the same structural form as the well-known linear regression model.

The power of neural network regression comes from constructing an architecture of one or more hidden layers to learn complex (nonlinear) relationships between the input features and the outcome variable. To demonstrate how a more complicated neural network generates a prediction, Figure 10.12 displays a neural network with two input variables, one hidden layer of three neurons, and one output variable. Figure 10.12 displays the system of neuron weights, bias terms, and activation functions that the neural network applies in a series of calculations. For every neuron in a layer, there is a weight associated with its connection to each neuron in the next layer. In Figure 10.12, the connection between neuron 1 and neuron 3 has weight w_{13} In addition, there is a bias term associated with each neuron in the hidden layer and output layer. In Figure 10.12, the bias term b_4 is associated with neuron 4. At each neuron in the hidden layer and output layer, an activation function is applied to the weighted sum of the inputs and corresponding bias term to generate an output value for the neuron. In Figure 10.12, the output of neuron 5 is produced as $\hat{x}_5 = f(w_{15}x_1 + w_{25}x_2 + b_5)$. The output of the neuron in the output layer corresponds to the predicted value of the outcome variable, $\hat{y} = f(w_{36}\hat{x}_3 + w_{46}\hat{x}_4 + w_{56}\hat{x}_6 + b_6)$.

Now let's consider applying the neural network with the structure in Figure 10.12 to predict the sales prices of pre-owned cars in the Palanquin example using information about how old the car is (Age) and how many miles the car has been driven (Odom). To compute a prediction using the neural network, we need values of the neuron weights (w_{13}, w_{14}, w_{15}, w_{23},…), the bias terms (b_3, b_4, b_5, b_6), and the form of the activation functions (f) at the neurons in the hidden layer and output layer. In the next subsection, we describe how a neural network iteratively learns the good values of the neuron weights and bias features, but for now we will use the initial values shown in Figure 10.13.

Consider a neural network to predict the price of a nine-year old car with an odometer reading of 95,000 miles. For a neural network, it is common to scale the value of each

Figure 10.12 Neuron Weights, Bias Terms, and Activation Functions of a Neural Network



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

input feature x to be between 0 and 1. **Min-max scaling** of a variable x is achieved with equation (10.1).

$$x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{10.1}$$

where x_{min} and x_{max} refer to the minimum and maximum observations, respectively, for the variable x. For example, suppose that the minimum and maximum values of Age in the training set used to estimate the parameters of the neural network are 4 and 20 years. Then the scaled value of Age for a nine-year old car is:

$$\text{Age}_{\text{scaled}} = \frac{\text{Age} - \text{Age}_{\text{min}}}{\text{Age}_{\text{max}} - \text{Age}_{\text{min}}} = \frac{9-4}{20-4} = 0.3125$$

Similarly, suppose that the minimum and maximum values of Odom in the training set used to estimate the parameters of the neural network are 50,500 and 250,500 miles. Then the scaled value of Odom = 95,000 is:

$$\text{Odom}_{\text{scaled}} = \frac{\text{Odom} - \text{Odom}_{\text{min}}}{\text{Odom}_{\text{max}} - \text{Odom}_{\text{min}}} = \frac{95{,}000 - 50{,}500}{250{,}500 - 50{,}500} = 0.2225$$

As shown in Figure 10.13, the values of $\text{Age}_{\text{scaled}} = 0.3125$ and $\text{Odom}_{\text{scaled}} = 0.2225$ are the values for the input layer.

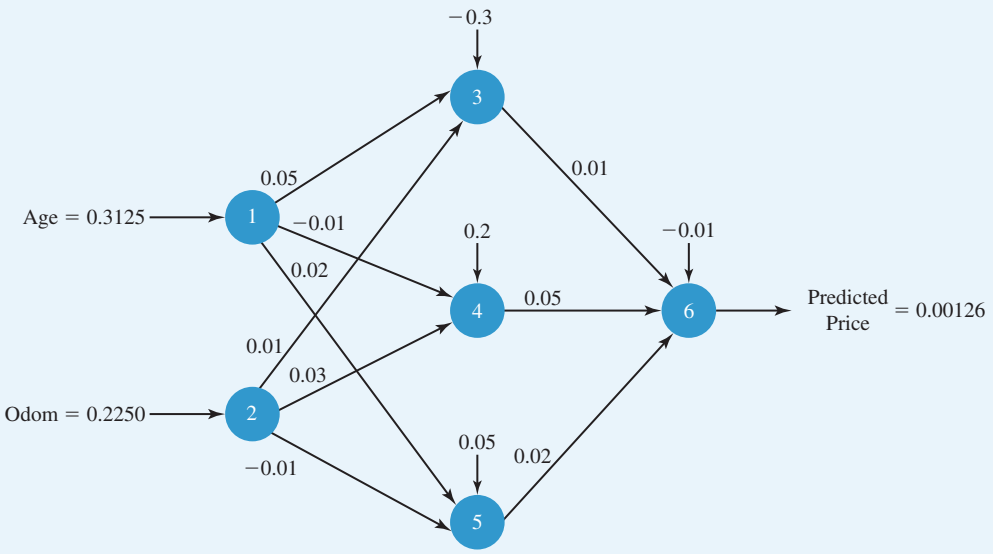
Given the values of the input features, the output values of the neurons in the next layer can be computed for a specified activation function. A common activation function in the hidden layers of a neural network is called the **rectified linear unit (ReLu) function**. The ReLu function is

$$f^{\text{ReLu}}(x) = \max\{0, x\}$$

Using the ReLu function, we obtain the following output values for the neurons in the hidden layer of Figure 10.13:

$$\begin{aligned}\hat{x}_3 = \max\{0, b_3 + w_{13}x_1 + w_{23}x_2\} &= \max\{0, -0.3 + (0.05)(0.3125) + (0.01)(0.225)\} \\ &= \max\{0, -0.282\} = 0\end{aligned}$$

Figure 10.13 Initial Prediction for Pre-owned Car Price from Neural Network



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$$\hat{x}_4 = \max\{0, b_4 + w_{14}x_1 + w_{24}x_2\} = \max\{0, 0.2 + (-0.01)(0.3125) + (0.03)(0.225)\} = \max\{0, 0.204\} = 0.204$$

$$\hat{x}_5 = \max\{0, b_5 + w_{15}x_1 + w_{25}x_2\} = \max\{0, 0.05 + (0.02)(0.3125) - (0.01)(0.225)\} = \max\{0, 0.054\} = 0.054$$

The output values of the three neurons in the hidden layer then become inputs to the computation of the neuron in the output layer. For a regression task, a common activation function for the output layer is the identity function, $f(x) = x$.

$$\hat{y} = b_6 + w_{36}\hat{x}_3 + w_{46}\hat{x}_4 + w_{56} = -0.01 + (0.01)(0) + (0.05)(0.204) + (0.02)(0.054) = 0.00126$$

Thus, for Age = 0.3125 and Odom = 0.2250, this neural network produces a predicted sales price of 0.00126. For a neural network regression, it is common to standardize the outcome variable by subtracting its mean and dividing by its standard deviation. Suppose the mean price in the training set ($\bar{x}$) and the price standard deviation in the training set (s) are \$16,075 and \$14,166, respectively. Then, we can obtain the predicted price in its original units (\$) with the calculation

$$\text{Price (\$)} = \bar{x} + \text{Price}_{\text{stand}} \times s = 16{,}075 + (0.00126)(14{,}166) = 16{,}093$$

Thus, the neural network with these initial parameters predicts a price of \$16,093 for this nine-year old vehicle with 95,000 miles. Suppose that the actual sales price of this vehicle is \$14,000, or (14,000 − 16,075)/14,166 = −0.1465 in the standardized price units. Thus, the prediction of the neural network with these initial neuron weights and bias terms results in a prediction error of 14,000 − 16,093 = −2,093 in the original units (\$), and a prediction error of −0.1465 − 0.00126 = −0.1477 in standardized units. In the next section, we describe how the neural network iteratively updates the values of its weight and bias parameters based on the prediction error in an effort to improve its predictive accuracy.

How a Neural Network Learns

To help explain the process of how the parameters of a neural network are estimated, consider a training set of eight observations of pre-owned cars from the Palanquin Auto example shown in Table 10.18. The variable values for these eight observations have been scaled according to equation (10.1).

Table 10.18 Min-Max Scaled Training Set Observations

Car	Age (years)	Odom (miles)	$\text{Age}_{\text{scaled}}$	$\text{Odom}_{\text{scaled}}$	Price (\$)	$\text{Price}_{\text{stand}}$
1	9	95,000	0.3125	0.2225	14,000	−0.1465
2	8	50,500	0.2500	0.0000	15,000	−0.0759
3	7	166,380	0.1875	0.5794	22,950	0.4853
4	20	250,500	1.0000	1.0000	900	−1.0712
5	12	191,955	0.5000	0.7073	20,950	0.3441
6	13	204,561	0.5625	0.7703	2,500	−0.9583
7	14	181,000	0.6250	0.6525	7,300	−0.6194
8	4	62,000	0.0000	0.0575	45,000	2.0419
Max:	20	250,500		$\bar{x}$:	16,075	
Min:	4	50,500		s:	14,166	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The process of training a neural network begins with the initialization of values for neuron weights and bias terms. These initial parameters are typically set to small random values between −0.5 and 0.5 such as those shown in Figure 10.13.

A single iteration of the neural network training process consists of the following steps:

1. Predict the outcome value for each observation in a batch using the current parameter values.
2. Update the parameter values based on the prediction error on the current batch of training set observations.

The **batch size** is the number of training set observations whose outcome values are predicted before updating the parameter values. The batch size is a user-defined setting and can be as small as one (corresponding to the case in which the parameter values are updated after predicting a single observation) to the size of the training set (corresponding to the case in which the parameter values are updated after predicting all the observations in the training set).

Now we demonstrate the neural network training process using the training set in Table 10.18 and a batch size of four. The first training iteration considers observations 1 to 4. Starting with the initial parameter values in Figure 10.13, Table 10.19 contains the predicted prices for the first iteration.

In Table 10.19, the overall quality of the predictions for the batch of observations 1 to 4 is measured by the sum of the squared error (SSE). Recall that the neural network generates each predicted car price through a forward pass of prediction calculations from the input layer to the output layer. In contrast, the parameters of the neural network (neuron weights and bias terms) are updated through a backward pass of adjustment calculations from the output layer to the input layer. A method called **backpropagation** employs the chain rule[4] from calculus to estimate how much a change in each of the respective parameters (neuron weights, bias terms) affects the prediction error of a batch of training set observations. Data mining software implements backpropagation to determine amount to adjust each neuron weight and bias term. Let Δ_{ij} be the amount that the neuron weight w_{ij} is updated. Let Δ_i be the amount that the bias term b_i is updated. Then, the update formulas for a feature weight w_{ij} and a bias term b_i are given by

$$w_{ij}^{new} = w_{ij} + \Delta_{ij} \tag{10.2}$$

$$b_i^{new} = b_i + \Delta_i \tag{10.3}$$

After completing the update of all parameters, the next iteration begins by using the new values of the parameters to generate predictions for the next batch of observations (cars 5-8).

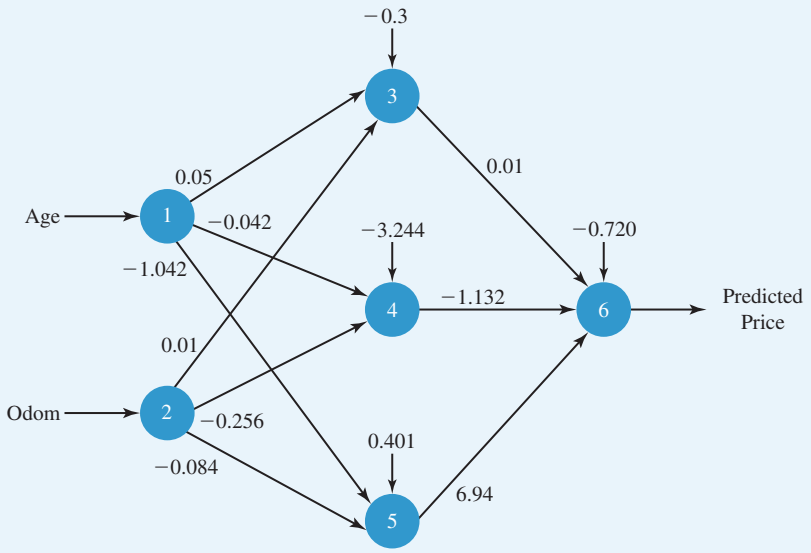
Table 10.19 Results of First Iteration of Neural Network Training

Car	Age_{scaled}	$Odom_{scaled}$	$Price_{stand}$ (y)	Predicted Price ($\hat{y}$)	$(y - \hat{y})^2$
1	0.3125	0.2225	−0.14648	0.00126	0.02183
2	0.2500	0.0000	−0.07589	0.00098	0.00591
3	0.1875	0.5794	0.48532	0.00173	0.23386
4	1.0000	1.0000	−1.07123	0.00220	1.15226
				SSE:	1.41385
				RMSE:	0.59453

[4]The chain rule is a formula for finding the derivative of the composition of two functions. The specifics of the chain rule, and its application to neural networks, are beyond the scope of this text. For more information see C. C. Aggarwal, *Neural Networks and Deep Learning*, Springer, 2018.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 10.14 Neuron Weights and Bias Terms After Several Iterations



Then, the sum of squared errors of the predictions on observations 5-8 is used to again adjust the parameters according to the update formulas in equation (10.2) and (10.3). These new values of the parameters are then used to generate predictions for the next batch of observations (cars 1-4 again). This process repeats until completing the specified number of iterations.

After several iterations of the parameter updating procedure, the neuron weights and bias terms in Figure 10.14 are obtained. The predicted prices in Table 10.20 are generated using the neuron weights and bias terms in Figure 10.14 using the ReLu activation function for the neurons in the hidden layer and the identity activation function for the neuron in the output layer. By comparing the RMSE = 0.5945 from the first iteration of neural network training in Table 10.19 to the RMSE = 0.4299 after several iterations of neural network training, we observe that the neural network's predictions have improved.

Table 10.20 Training Set Predictions for Neural Network with Updated Neuron Weights and Bias Terms

Car	Age_{scaled}	$Odom_{scaled}$	$Price_{stand}$	Predicted Price	Error
1	0.3125	0.2225	−0.14648	−0.32717	0.18069
2	0.2500	0.0000	−0.07589	0.25477	−0.33066
3	0.1875	0.5794	0.48532	0.36865	0.11667
4	1.0000	1.0000	−1.07123	−0.71978	−0.35145
5	0.5000	0.7073	0.34414	−0.71978	1.06392
6	0.5625	0.7703	−0.95829	−0.71978	−0.23851
7	0.6250	0.6525	−0.61945	−0.71978	0.10033
8	0.0000	0.0575	2.04187	2.02949	0.01239
				RMSE:	0.42985

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 10.21 Validation Set Predictions for Neural Network with Updated Neuron Weights and Bias Terms

Car	Age_{scaled}	$Odom_{scaled}$	$Price_{stand}$	Predicted Price	Error
9	0.3750	0.5675	–0.50297	–0.71978	0.21681
10	0.7500	0.0141	–0.50734	–0.71978	0.21243
11	0.3750	0.7563	0.03000	–0.71978	0.74978
12	0.5625	0.0310	–0.49944	–0.71978	0.22034
				RMSE:	0.41918

To evaluate the out-of-sample predictive accuracy of the neural network with the values of neuron weights and bias terms shown in Figure 10.14, we apply the neural network to the four validation set observations in Table 10.21.

Neural networks have the advantage of being able to capture complex relationships between the input features and the outcome variable. The disadvantage is that neural networks typically need more data (more observations) to effectively estimate the neuron weights and bias terms necessary to capture these complex relationships. Additionally, a neural network is not easily interpreted and often treated as a "black box" with model construction consisting of comprehensive testing of the many algorithm choices including the number of hidden layers, number of neurons in a hidden layer, activation functions, etc.

Notes + Comments

1. In addition to the feed forward neural network architecture discussed in this section in which information flows unidirectionally, there is another type of structure called a recurrent neural network. A recurrent neural network allows information to flow in both directions by introducing loops between the neurons of the network.
2. In addition to the ReLu function, other common functions used as activation functions in the hidden layers of neural networks for regression tasks include the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$, the hyperbolic tangent function $f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$, and the softplus function $f(x) = ln(1 + e^{x})$.
3. Scaling the input variables and outcome variable is an important pre-processing step in neural network regression because it facilitates the updating of the parameters as the neural network learns from the training data. Min-max scaling or standardizing a variable value by subtracting the mean and dividing by the standard deviation are common techniques. The best way to scale the variables may be dependent on the data, so experimenting with both ways is recommended.
4. The training and validation sets we used in this section to demonstrate neural networks are impractically small and only suitable for illustration of the process.

10.6 Feature Selection

In this chapter, we have described how to apply *k*-nearest neighbors, regression trees, and neural networks to a regression task. These three methods, along with linear regression, each use the information in the set of input variables $x_1, x_2, \ldots, x_q$ in a different manner to generate predictions for a quantitative outcome variable y. In this section, we discuss how to select the input variables to include in the construction of a predictive model, a process called **feature selection**. Feature selection can be particularly important when dealing with data sets with many variables because including all the variables in the predictive model can lead to inferior predictions due to the presence of redundant and irrelevant variables. Feature selection can reduce the chance of overfitting a model, improve

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

predictive performance if a good subset of variables is identified, and reduce the time required to train a model.

There are three types of feature selection methods: wrapper methods, filter methods, and embedded methods. The choice of feature selection method depends on the type of predictive method and the data.

Wrapper Methods

Given data on an outcome variable y and input variables $x_1, x_2, \ldots, x_m$, a **wrapper method** iteratively executes the following steps:

1. Select a subset of $q \leq m$ input variables.
2. Conduct the training and validation of a predictive model considering the q selected input variables.

A wrapper method repeats steps 1 and 2 for a specified number of iterations. The simplest wrapper method considers all possible subsets of variables, which is called the **best subsets method**. The best subsets method is guaranteed to identify the subset of input variables which results in the best out-of-sample predictive accuracy. However, the best subsets approach requires 2^m iterations of training and validation, which is impractical for data sets with many variables.

We discuss variable selection procedures guided by variables' p-values for linear regression models in Chapter 8. Backward elimination, forward selection, and stepwise selection discussed in Section 8.8 are examples of wrapper methods for linear regression models.

Rather than exhaustively considering all possible subsets of variables, wrapper methods are designed to try to intelligently select the next subset of variables to consider at an iteration. Some wrapper methods are tailored to a specific predictive method. For example, if considering a linear regression model, there are wrapper methods which consider the p-values of the variables to determine the next subset to evaluate.

Given enough time to evaluate different subsets of variables, a wrapper method often identifies a good (if not the best) subset of variables. However, wrapper methods may require a substantial amount of time to search among possible subsets of input variables when m is large, and unless the method considers all 2^m subsets, it cannot guarantee finding the set of input variables $x_1, x_2, \ldots, x_q$ that achieves the best out-of-sample predictive accuracy.

Filter Methods

Given data on an outcome variable y and input variables $x_1, x_2, \ldots, x_m$, a **filter method** uses statistical analysis to screen out the input variables that seem unlikely to be helpful in predicting the outcome variable. To identify variables that are potentially irrelevant, filter methods consider the statistical analysis of the relationship between the quantitative outcome variable y and an input variable x.

If the input variable x under consideration is quantitative, then a scatter chart can be constructed to visualize its relationship with the quantitative outcome variable y. In addition, it is common to measure the relationship between a quantitative input variable x and the quantitative outcome variable y by computing the **correlation coefficient** of x and y using the training data. If the correlation coefficient of x and y is near zero, then the input variable x may be irrelevant and possibly removed. However, how close to zero a correlation coefficient must be to reliably identify an irrelevant quantitative input variable varies, and so removing a variable is still up to the analyst's discretion.

We introduce the concept of correlation in Chapter 2.

If the input variable x under consideration is categorical, then its relationship with the quantitative outcome variable y can be visualized by constructing boxplots of y for each category of x. If the resulting boxplots for each category are very similar, then the input variable x may not add information and possibly can be removed. In addition, it is common to measure the relationship between a categorical input variable x and a quantitative outcome variable y by computing a **confidence interval** on the mean of y for each categorical value of x. If the difference between the sample mean values of y for the different categories of x is not statistically significant, then the input variable x may be irrelevant and possibly removed. However, the level of statistical significance that reliably identifies an irrelevant categorical input variable varies, and so removing a variable is still up to the analyst's discretion.

We introduce confidence intervals in Chapter 7.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The advantage of filter methods is that they are relatively easy and fast to execute. A disadvantage is that the statistical measures employed in filter methods do not explicitly consider the performance of the predictive model and require the analyst to determine the thresholds for determining irrelevance. Therefore, filter methods may fail to identify the best collection of input variables.

Embedded Methods

Given data on an outcome variable y and input variables $x_1, x_2, \ldots, x_m$, an **embedded method** simultaneously selects the features to consider in the model as it trains the model. An embedded method is tailored to the predictive model that is being considered.

One example of a technique used for embedded feature selection is regularization. **Regularization** is a general technique to avoid overfitting a predictive model by controlling the model complexity. In this section, we describe penalty-based regularization methods.

For example, in linear regression, model complexity is determined by the number of input variables. Recall that for linear regression, the n observations in the training data are used to estimate the coefficients $b_0, b_1, \ldots, b_q$ to generate predictions $\hat{y} = b_0 + b_1x_1 + \cdots b_qx_q$ that minimize sum of squared error:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{10.4}$$

A regularization approach to linear regression is to add a penalty to equation (10.4) so that the coefficients $b_0, b_1, \ldots, b_q$ are estimated to minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{q}\left|b_j\right|^p \tag{10.5}$$

In equation (10.5), λ is a tuning parameter and p is a parameter that is commonly set to either 1 or 2. For a value of $p = 1$, equation (10.5) is called **lasso regularization**. For a value of $p = 2$, equation (10.5) is called **ridge regularization**. Because the coefficients $b_1, \ldots, b_q$ are estimated to minimize equation (10.5), adding a regularization penalty tends to make the coefficient values closer to zero. The value of λ in equation (10.5) that results in the values of $b_0, b_1, \ldots, b_q$ that achieve the best out-of-sample predictive accuracy is determined by iteratively varying λ in a series of cross-validation experiments. For this best value of λ, the values of the coefficients $b_1, \ldots, b_q$ are examined. For linear regression, lasso regularization serves as embedded feature selection as it often results in coefficient values equal to zero for one or more variables, suggesting that these variables can be dropped from the model.

Similar to the manner in which regularized linear regression estimates its coefficients to minimize equation (10.5), regularization can be embedded in the training of a neural network. For a regression task, a regularized neural network estimates its parameters (neuron weights and bias terms) to minimize an equation analogous to equation (10.5), in which the regression coefficients are replaced with neural network parameters. Such a regularization approach helps prevent the neural network from overfitting the training data by forcing some of the weight values closer to zero.

Another example of embedded feature selection occurs in the construction of a regression tree. Recall that as a regression tree is being constructed on training data, the tree logic splits a node of observations by considering each input variable to identify the variable which results in the subsets with the largest decrease in variance of the outcome variable. The presence of extraneous input variables generally does not affect the construction of the classification tree. This is because the extraneous variable should not be selected as the basis of a splitting rule because it will not decrease the impurity of the outcome variable in the resulting subsets.

While feature selection is embedded in the construction logic of a regression tree, regularization can be useful to prune regression trees. In a regression tree, model complexity is determined by the number of rules. Recall that for a regression tree, the n observations in the training data are used to a set of rules that divide the observations into groups and then

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

these rules are used to generate predictions $\hat{y}$. A regularization approach to constructing a regression tree is to determine the set of rules using the training set to minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda|T| \tag{10.6}$$

In equation (10.6), λ is a complexity parameter and $|T|$ corresponds to the number of leaf nodes in the tree. The regularization penalty in equation (10.6) makes trees with more branches more costly, so increasing the value of λ implicitly prunes the regression tree. The value of λ in equation (10.6) that results in the tree that achieves the best out-of-sample predictive accuracy is determined by iteratively varying λ in a series of cross-validation experiments.

Notes + Comments

1. Feature selection can be critical when performing *k*-nearest neighbors regression. When computing similarity on the basis of too many features, the distance between observations becomes less distinct with each additional feature and thus less helpful in making predictions.
2. Although regression trees have embedded feature selection, the appropriate representation of the variables (part of the data preparation phase) is important as regression trees are susceptible to poorly coded categorical variables (particularly categorical variables with many values).
3. Feature selection can improve neural network regression by reducing the number of parameters (feature weights and constant factors) that need to be estimated from the training data.

Summary

In this chapter, we introduced data mining concepts and techniques for regression tasks which involve a quantitative outcome variable. We began by describing mean absolute error (MAE) and root mean squared error (RMSE) which are common performance measures used to assess prediction models for regression tasks. Using linear regression, we demonstrated the experimental process of assessing out-of-sample predictive accuracy of a predictive model with either the static holdout method or *k*-fold cross-validation. Then, we described several different predictive methods for a regression task. First, we discussed the use of *k*-nearest neighbors for a regression task. Then, we described regression trees and how combining many regression trees with an ensemble method generally results in more accurate predictions than a single regression tree. The last predictive method for a regression task we considered was neural networks. We showed how a neural network is a generalization of linear regression model and described the process of training a neural network and making predictions with it. We concluded the chapter with a discussion of methods for selecting features for predictive models and the use of regularization techniques to avoid overfitting.

Glossary

Activation function A function in a neuron of a neural network that produces an output from a neuron based on the inputs into a neuron.
Backpropagation A process by which the parameters of the neural network (neuron weights and bias terms) are updated through a backward pass of adjustment calculations from the output layer to the input layer.
Bagging An ensemble method that generates a committee of models, each based on a slightly different training set, and makes predictions based on the average prediction of the set of models.
Batch size The number of training set observations whose outcome values are predicted before updating the parameter values in a neural network.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Best subsets method A wrapper method for selecting input variables for a predictive model that considers all possible 2^m subsets of variables (where m is the total number of variables under consideration).
Bias term A parameter value that amplifies or dampens the input into a neuron of a neural network.
Boosting An ensemble method that iteratively adapts the construction of the jth training set based on the prediction error of the learners constructed on the previous $j - 1$ training sets.
Correlation coefficient A standardized measure of linear association between two variables that takes on values between −1 and +1.
Confidence interval An estimate of a population parameter (e.g., mean) that provides an interval believed to contain the value of the parameter at some level of confidence.
Embedded method A method that selects features for a predictive model as part of the model training and validation process.
Ensemble method A predictive data mining approach in which a committee of individual prediction models are generated and a prediction is made by combining these individual predictions.
Feature An input variable used for predicting values of an outcome variable.
Feature selection The process of determining the input variables to include in the construction of a predictive model.
Feed forward neural network A neural network in which the prediction calculations are processed in only one direction.
Filter method A method for selecting features for a predictive model by analyzing the statistical relationships of the features and screening out variables likely to be irrelevant.
Hidden layer A layer of neurons in a neural network between the input layer and the output layer.
Input layer The first layer of neurons in a neural network through which the values of the input features enter.
***k*-fold cross-validation** A procedure to train and validate models in which the observations to be used to train and validate the model are repeatedly randomly divided into k subsets called folds. In each iteration, one fold is designated as the validation set and the remaining $k - 1$ folds are designated as the training set. The results of the iterations are then combined and evaluated.
***k*-nearest neighbors (*k*-NN)** A data mining method that predicts an observation i's outcome value based on the k observations most similar to observation i with respect to the input variables.
Lasso regularization Regularization in which the penalty term considers the absolute value function.
Lazy learner A predictive model that performs no calculations on the training set prior to being supplied an observation to predict.
Leaf node A node in a tree that is not further split into more nodes.
Leave-one-out cross-validation A special case of k-fold cross validation for which the number of folds equals the number of observations in the combined training and validation data.
Mean absolute error (MAE) The average of the absolute values of the prediction errors.
Min-max scaling The scaling of a variable so that its values lie in the range between 0 and 1.
Neural network A predictive data mining model whose structure is motivated by the biological functioning of the brain.
Neuron A node in a neural network.
Neuron weight A parameter value representing the strength of a neuron's output signal to a neuron in the next layer of a neural network.
Out-of-bag estimation A measure of estimating the predictive performance of a bagging ensemble of m models (without a separate validation set) by leveraging the concept that the training of each model is only based on approximately 63.2% of the original observations (due to sampling with replacement).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Output layer The final layer of neurons in a neural network through which the predicted outcome values are obtained.
Overfitting A situation in which a model explains random patterns in the data on which it is trained rather than just the generalizable relationships, resulting in a model with training-set performance that greatly exceeds its performance on holdout data.
Perceptron A neural network consisting only of an input layer and output layer (no hidden layers).
Prediction error The difference between the observed value and the predicted value.
Pruning The process of obtaining a tree with fewer leaf nodes than a full tree to prevent overfitting on training data.
Random forest A variant of the bagging ensemble method that generates a committee of trees based on different random samples but restricts each individual tree to a limited number of randomly selected features (variables).
Rectified linear unit (ReLu) function A function $f(x)$ that is equal to 0 if $x < 0$ and equal to x if $x \geq 0$.
Regression A prediction task in which the variable to be predicted is quantitative.
Regression tree A tree that predicts values of a continuous outcome variable by splitting observations into groups via a sequence of hierarchical rules on the input variables.
Regularization A penalty-based method to avoid overfitting a predictive model by controlling model complexity.
Ridge regularization Regularization in which the penalty term considers a quadratic function.
Root mean squared error (RMSE) A performance measure computed as the square root of the average squared prediction error.
Root node The top node in a tree containing all the observations in the training set.
Stump A tree with a only a single rule.
Supervised learning Category of data mining techniques in which an algorithm learns how to classify or estimate an outcome variable of interest.
Test set Data set used to compute unbiased estimate of final predictive model's performance.
Training set Data used to build candidate predictive models.
Unstable learner A prediction model for which small changes in the training set cause a model's predictions to fluctuate substantially.
Validation set Data used to evaluate candidate predictive models.
Weak learner A predictive model whose individual predictions are only slightly better than random guessing.
Wrapper method A method for selecting features for a predictive model that iteratively tests different subsets of features.

Problems

Conceptual

Problems 1 through 14 do not require the use of data mining software and focus on knowledge of concepts and basic calculations.

1. **Performance measures for Regression Tasks.** Which of the following statements is false? LO 1
   a. RMSE and MAE often agree on which regression model is the most accurate.
   b. RMSE and MAE both indicate when the predicted values are consistently larger than the actual values.
   c. RMSE will penalize a prediction method for a large prediction error more than MAE.
   d. RMSE and MAE both measure error in the units of the outcome variable.
2. **Assessing Predictive Models.** What is a benefit of k-fold cross-validation over the static holdout method? LO 2
   a. The k-fold cross-validation method is computationally faster.
   b. The k-fold cross-validation method is easier to implement when the standardization or normalization of variables is required.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. For $k = 1$, the k-fold cross-validation method is the same as the static holdout method.
d. The k-fold cross-validation method results in an out-of-sample prediction for every observation in the data.

3. **Regression with *k*-Nearest Neighbors**. Why is the k-nearest neighbors regression called a lazy learner? LO 3
   a. Because k-nearest neighbors does not have embedded feature selection.
   b. Because the k-nearest neighbors regression requires the standardization or normalization of the variables.
   c. Because the k-nearest neighbors regression does not use training data to estimate model parameters or determine variable splitting rules, but instead directly uses the average of the k most similar training set observations as its predicted value for a new observation.
   d. Because the k-nearest neighbors method does not require the use of a test set.

4. **Executing *k*-Nearest Neighbors Regression.** Which of the following statements about k-nearest neighbors regression is false?
   a. The choice of input features may have an impact on predictive performance.
   b. Variables do not need to be standardized.
   c. If too large of a value of k is selected, prediction error will be large due to underfitting.
   d. If too small of a value of k is selected, prediction error will be large due to overfitting.

5. **Regression Tree Splits**. When splitting a node of a regression tree into two subsets, how are the variable and its split value determined? LO 4
   a. At each step, the variable and its split value that are selected result in the largest decrease in the variance of the outcome variable.
   b. The variable that has a split value with the smallest regularization penalty is selected.
   c. At each step, the selected variable and split value result in the most balanced split of the observations in the parent node into two children nodes.
   d. At each step, a variable is randomly selected and then the split value is determined to be the one that results in the largest decrease in the variance of the outcome variable between the parent region and the two new children regions.

6. **Pruning Regression Trees**. Why is it generally a good idea to prune a full regression tree? LO 4
   a. A full regression tree is too large to visualize.
   b. A full regression tree typically overfits the training data and pruning the tree improves the predictive performance on validation data.
   c. Pruning a tree by applying restrictions on how much a variable splitting rule can divide up a node of observations decreases the training set variance of the outcome variable in the resulting leaf nodes.
   d. Pruning a tree results in a weak, unstable learner that is better for out-of-sample predictions.

7. **Ensemble Methods**. Which of the following statements is true? LO 5
   a. A bagging ensemble is more computationally expensive than a boosting ensemble.
   b. An ensemble method requires a collection of strong, stable individual learners to be effective.
   c. Random forests try to decorrelate their constituent regression trees by basing them on slightly different training sets and by randomly restricting the variables they can use to generate rules.
   d. A boosting ensemble can be constructed in a parallel because the individual models do not rely on each other.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

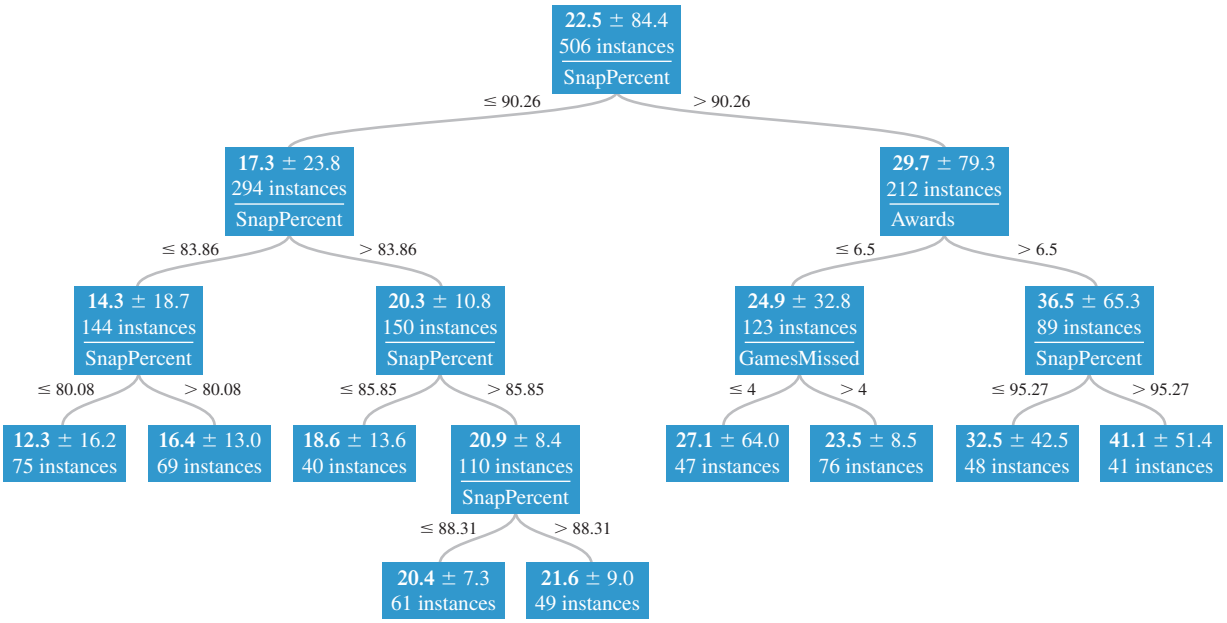
8. **Comparing Ensemble Types.** Which of the following statements about ensemble methods is false? LO 5
   a. In a boosting ensemble, the training data for an individual model depends on the predictions of the previously constructed individual models.
   b. A bagging ensemble generates a composite prediction by averaging individual predictions from a collection of unstable learners trained on slightly different sets of training data.
   c. A random forest is a type of bagging ensemble with an additional mechanism that attempts to make the individual trees more independent of each other.
   d. In boosting, individual models can be trained independently of each other in parallel.
9. **Neural Network Architecture**. Which of the following statements about the architecture of a neural network is false? LO 6
   a. The output layer of a neural network for a regression task commonly consists of only one neuron.
   b. The number of neurons in the input layer typically corresponds to the number of variables under consideration.
   c. The activation function in the output layer of a neural network for a regression task is commonly the identity function.
   d. There is no need to scale the values of the input variables.
10. **Neural Network Properties**. Which of the following statements about neural networks is false? LO 6
   a. It can be very easy to overfit a neural network due to number of weights to be estimated and the number of ways these weights are combined.
   b. A neuron's output in a neural network is generated by applying an activation function to the weighted combination of the values from the neurons from the previous layer.
   c. A neural network with no hidden layers and the identity activation function for the output layer results in a prediction model analogous to linear regression.
   d. A neural network is a highly flexible predictive supervised learning method that does not require the estimation of any parameters.
11. **Neural Network Training**. Which of the following statements about the training of a neural network is false? LO 6
   a. The training of a neural network is an iterative process of predicting an observation or batch of observations and then updating neuron weights based on a function of the prediction error.
   b. The performance of a neural network does not depend on the choice of activation function used for the hidden layers.
   c. To address the overfitting of a neural network, a regularization approach can be applied to penalize the magnitude of the weights placed on neurons' output.
   d. The size and number of hidden layers is determined experimentally by analyzing the impact on predictive performance in a cross-validation experiment.
12. **Feature Selection.** Which of the following statements is true? LO 7
   a. Filter methods are guaranteed to determine the best set of variables to include in a model.
   b. The best subsets method is an example of a wrapper method.
   c. Regularization is used with *k*-nearest neighbors regression to prevent too large of a value of *k* being selected.
   d. The *k*-nearest neighbors algorithm has an embedded feature selection procedure.
13. **Regularization.** Which of the following statements is true? LO 8
   a. Lasso regularization and ridge regularization are different methods for standardizing variables.
   b. Lasso regularization may prevent overfitting of a linear regression model by driving the coefficients of some variables to zero.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Regularization is not useful in the training of a regression tree because it has embedded feature selection.
d. Regularization is not helpful in the training of neural networks when there is sufficient data for the neural network.

14. **Athlete Contract Negotiations.** Casey Deesel is a sports agent negotiating a contract for Titus Johnston, an athlete in the National Football League (NFL). An important aspect of any NFL contract is the amount of guaranteed money over the life of the contract. Casey has gathered data on NFL athletes who have recently signed new contracts. Each observation (NFL athlete) includes values for percentage of the team's plays that the athlete is on the field (SnapPercent), the number of awards an athlete has received recognizing on-field performance (Awards), the number of games the athlete has missed due to injury (GamesMissed), and millions of dollars of guaranteed money in the athlete's most recent contract (Money, dependent variable).

Casey has trained a full regression tree on 304 observations and then used the validation set to prune the tree to obtain a pruned tree. The pruned tree (as applied to the data on 506 NFL athletes) follows. **LO 4**



a. Titus Johnston's variable values are: SnapPercent = 96, Awards = 6, and GamesMissed = 3. How much guaranteed money does the regression tree predict that a player with Titus Johnson's profile should earn in their contract?
b. Casey feels that Titus was denied an additional award in the past season due to some questionable voting by some sports media. If Titus had won this additional award, how much additional guaranteed money would the regression tree predict for Titus versus the prediction in part (a)?

Problems 15 through 29 and the Case Problem require the use of data mining software (such as R or Orange). Guidelines for solving these problems with R and Orange are provided in the respective appendixes available in the Cengage eBook.

Software Application

15. **Credit Scores: Regression Tree.** A consumer advocacy agency, Equitable Ernest, is interested in providing a service that allows an individual to estimate their own credit score (a continuous measure used by banks, insurance companies, and other businesses when granting loans, quoting premiums, and issuing credit). The file *creditscore* contains data from several thousand individuals. The variables in these data are listed in the following table. **LO 1, 2, 4**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file* creditscore

Variable	Description
BureauInquiries	number of inquiries about an individual's credit
CreditUsage	percent of an individual's credit used
TotalCredit	total amount of credit available to individual
CollectedReports	number of times an unpaid bill was reported to collection agency
MissedPayments	number of missed payments
HomeOwner	1 if individual is homeowner. 0 if not
CreditAge	average age of individual's credit
TimeOnJob	how long the individual has been continuously employed
CreditScore	score between 300 and 850 with larger number representing increased credit worthiness

Predict the individuals' credit scores using an individual regression tree. Use 100% of the data for training and validation and set aside no data as a test set.

a. Train a full regression tree and report its RMSE from a validation experiment.
b. Train a pruned regression tree and report its RMSE from a validation experiment.
c. Compare the RMSE from the full regression tree in part (a) to the RMSE from the pruned regression tree in part (b). Explain the difference.
d. Consider an individual with 5 credit bureau inquiries, that has used 10% of their available credit, has $14,500 of total available credit, has no collection reports or missed payments, is a homeowner, has an average credit age of 6.5 years, and has worked continuously for the past 5 years. Using the pruned tree from part (b), what is the predicted credit score for this individual?

16. **Credit Scores: Random Forest.** A consumer advocacy agency, Equitable Ernest, is interested in providing a service that allows an individual to estimate their own credit score (a continuous measure used by banks, insurance companies, and other businesses when granting loans, quoting premiums, and issuing credit). The file *creditscore* contains data from several thousand individuals. The variables in these data are listed in Problem 15. Predict the individuals' credit scores using a random forest of regression trees. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 5**

DATA *file* creditscore

a. Determine the random forest configuration that minimizes the RMSE in a validation procedure.
b. For the best-performing random forest identified in part (a), what is the RMSE on the test set?

17. **Credit Scores: *k*-Nearest Neighbors.** A consumer advocacy agency, Equitable Ernest, is interested in providing a service that allows an individual to estimate their own credit score (a continuous measure used by banks, insurance companies, and other businesses when granting loans, quoting premiums, and issuing credit). The file *creditscore* contains data from several thousand individuals. The variables in these data are listed in Problem 15. Predict the individuals' credit scores using a *k*-nearest neighbors. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 3**

DATA *file* creditscore

a. Based on all the input variables, determine the value of *k* that minimizes the RMSE in a validation procedure.
b. Experiment with different subsets of variables as input features and re-calibrate the value of *k* to minimize the RMSE. How does this *k*-nearest neighbors model compare to the model obtained in part (a)?
c. For the best-performing *k*-nearest neighbors model in the validation procedure, what is the RMSE on the test set?

18. **Credit Scores: Neural Networks.** A consumer advocacy agency, Equitable Ernest, is interested in providing a service that allows an individual to estimate their own credit

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file* creditscore

score (a continuous measure used by banks, insurance companies, and other businesses when granting loans, quoting premiums, and issuing credit). The file *creditscore* contains data from several thousand individuals. The variables in these data are listed in Problem 15. Predict the individuals' credit scores using a neural network. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 6**

a. Determine the neural network configuration that minimizes the RMSE in a validation procedure.
b. For the best-performing neural network identified in part (a), what is the RMSE on the test set?

19. **Credit Scores: Regularized Linear Regression.** A consumer advocacy agency, Equitable Ernest, is interested in providing a service that allows an individual to estimate their own credit score (a continuous measure used by banks, insurance companies, and other businesses when granting loans, quoting premiums, and issuing credit). The file *creditscore* contains data from several thousand individuals. The variables in these data are listed in Problem 15. Predict the individuals' credit scores using a linear regression model with lasso regularization. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 8**

DATA *file* creditscore

a. Determine the value of the lasso regularization penalty that minimizes the RMSE in a validation procedure.
b. For the best-performing linear regression model identified in part (a), what is the RMSE on the test set?
c. What are the coefficient values for the variables? Does the lasso regularization remove any variables from the model?

20. **Real Estate: Regression Tree.** CD Real Estate specializes in residential real estate services in the state of California. To complement the experience and local market knowledge of its licensed realtors, CD Real Estate wants to develop an analytical tool to predict the value of real estate. The file *calireal* contains data on some census tracts in California. The variables in these data are listed in the following table. **LO 1, 2, 4**

DATA *file* calireal

Variable	Description
Longitude	Longitude in center of tract
Latitude	Latitude in center of tract
Age	Median age of housing
Rooms	Total number of rooms
Bedrooms	Total number of bedrooms
Population	Population of tract
Households	Number of households
Income	Median income
Value	Median house value of tract (in $1,000s)

Predict the median house value using an individual regression tree. Set aside 50% of the data as a test set and use 50% of the data for training and validation.

a. Train a full regression tree and report its RMSE from a validation experiment.
b. Train a pruned regression tree and report its RMSE from a validation experiment.
c. Compare the RMSE from the full regression tree in part (a) to the RMSE from the pruned regression tree in part (b). Explain the difference.
d. Compute the RMSE of the pruned regression tree from part (b) on the test set.
e. Consider a tract with 5 credit bureau inquiries, that has a Longitude = −117.9, Longitude = 33.64, Age = 36, Rooms = 2,107, Bedrooms = 357, Population = 850, Households = 348, and Income = 5.0532. Using the pruned tree settings from part (b), what is the predicted median house value for this tract?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

21. **Real Estate: Random Forest.** CD Real Estate specializes in residential real estate services in the state of California. To complement the experience and local market knowledge of its licensed realtors, CD Real Estate wants to develop an analytical tool to predict the value of real estate. The file *calireal* contains data on some census tracts in California. The variables in these data are listed in Problem 20. Predict the tracts' median house values using a random forest. Set aside 50% of the data as a test set and use 50% of the data for training and validation. **LO 1, 2, 5**
    a. Determine the random forest configuration that minimizes the RMSE in a validation procedure.
    b. For the best-performing random forest identified in part (a), what is the RMSE on the test set?

DATA *file* calireal

22. **Real Estate: *k*-Nearest Neighbors.** CD Real Estate specializes in residential real estate services in the state of California. To complement the experience and local market knowledge of its licensed realtors, CD Real Estate wants to develop an analytical tool to predict the value of real estate. The file *calireal* contains data on some census tracts in California. The variables in these data are listed in Problem 20. Predict the individuals' credit scores using a *k*-nearest neighbors. Set aside 50% of the data as a test set and use 50% of the data for training and validation. **LO 1, 2, 3**
    a. Based on all the input variables, determine the value of *k* that minimizes the RMSE in a validation procedure.
    b. Experiment with different subsets of variables as input features and re-calibrate the value of *k* to minimize the RMSE. How does this *k*-nearest neighbors model compare to the model obtained in part (a)?
    c. For the best-performing *k*-nearest neighbors model in the validation procedure, what is the RMSE on the test set?

DATA *file* calireal

23. **Real Estate: Neural Networks.** CD Real Estate specializes in residential real estate services in the state of California. To complement the experience and local market knowledge of its licensed realtors, CD Real Estate wants to develop an analytical tool to predict the value of real estate. The file *calireal* contains data on some census tracts in California. The variables in these data are listed in Problem 20. Predict the individuals' credit scores using a neural network. Set aside 50% of the data as a test set and use 50% of the data for training and validation. **LO 1, 2, 6**
    a. Determine the neural network configuration that minimizes the RMSE in a validation procedure.
    b. For the best-performing neural network identified in part (a), what is the RMSE on the test set?

DATA *file* calireal

24. **Real Estate: Regularized Linear Regression.** CD Real Estate specializes in residential real estate services in the state of California. To complement the experience and local market knowledge of its licensed realtors, CD Real Estate wants to develop an analytical tool to predict the value of real estate. The file *calireal* contains data on some census tracts in California. The variables in these data are listed in Problem 20. Predict the individuals' credit scores using a linear regression model with lasso regularization. Set aside 50% of the data as a test set and use 50% of the data for training and validation. **LO 1, 2, 8**
    a. Determine the value of the lasso regularization penalty that minimizes the RMSE in a validation procedure.
    b. For the best-performing linear regression model identified in part (a), what is the RMSE on the test set?
    c. What are the coefficient values for the variables? Does the lasso regularization remove any variables from the model?

DATA *file* calireal

25. **Pre-Owned Car Prices: Regression Tree.** Palanquin Auto is a national pre-owned car retailer that offers an online marketplace for customers interested in purchasing a pre-owned automobile. In an effort to reduce price negotiation, Palanquin collects data on

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file* palanquin

automobile sales in an effort to set fair, competitive prices. The file *palanquin* contains data for the variables listed in the following table. **LO 1, 2, 4**

Variable	Description
Price	Sales price of vehicle
Year	Model year of vehicle
Manufacturer	Name of company manufacturing vehicle (categorical)
Model	Model name of vehicle (categorical)
Condition	Categorical variable describing condition of vehicle (fair, good, excellent, like new, new)
Cylinders	Number of cylinders in engine
Fuel	Type of fuel used by vehicle (diesel, electric, gas, hybrid)
Odometer	Number of miles on vehicle's odometer
Drive	Categorical variable indicating drivetrain (four wheel drive, front wheel drive, rear wheel drive)
Type	Categorical variable indicating
Paint	Color of vehicle's paint
State	State of the United States in which vehicle was sold

Predict the pre-owned car prices using an individual regression tree. Set aside 50% of the data as a test set and use 50% of the data for training and validation. Categorical input variables with many possible categories can pose computational difficulties for regression trees, so exclude any categorical variable with more than 16 categorical values from consideration as an input variable.

a. Train a full regression tree and report its RMSE from a validation experiment.
b. Train a pruned regression tree and report its RMSE from a validation experiment.
c. Compare the RMSE from the full regression tree in part (a) to the RMSE from the pruned regression tree in part (b). Explain the difference.
d. Compute the RMSE of the pruned regression tree from part (b) on the test set.
e. Using the pruned tree settings from part (b), predict the sales price of a vehicle with the following variable values:

Year	2016
Manufacturer	Dodge
Model	Charger
Condition	like new
Cylinders	6
Fuel	gas
Odometer	62,000
Drive	rwd
Type	sedan
Paint	silver
State	Alabama

26. **Pre-Owned Car Prices: Random Forest.** Palanquin Auto is a national pre-owned car retailer that offers an online marketplace for customers interested in purchasing a pre-owned automobile. In an effort to reduce price negotiation, Palanquin collects data on automobile sales in an effort to set fair, competitive prices. The file *palanquin* contains data for the variables listed in Problem 25. Predict the pre-owned car prices using a random forest. Set aside 50% of the data as a test set and use 50% of the data for training and validation. Categorical input variables with many possible categories can pose

DATA *file* palanquin

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

computational difficulties for tree-based methods, so exclude any categorical variable with more than 16 categorical values from consideration as an input variable. **LO 1, 2, 5**

a. Determine the random forest configuration that minimizes the RMSE in a validation procedure.
b. For the best-performing random forest identified in part (a), what is the RMSE on the test set?

27. **Pre-Owned Car Prices: *k*-Nearest Neighbors.** Palanquin Auto is a national used car retailer that offers an online marketplace for customers interested in purchasing a pre-owned automobile. In an effort to reduce price negotiation, Palanquin collects data on automobile sales in an effort to set fair, competitive prices. The file *palanquin* contains data for the variables listed in Problem 25. Predict the pre-owned car prices using *k*-nearest neighbors. Set aside 50% of the data as a test set and use 50% of the data for training and validation. **LO 1, 2, 3**

DATA *file* palanquin

a. Based on all the input variables, determine the value of *k* that minimizes the RMSE in a validation procedure.
b. Experiment with different subsets of variables as input features and re-calibrate the value of *k* to minimize the RMSE. How does this *k*-nearest neighbors model compare to the model obtained in part (a)?
c. For the best-performing *k*-nearest neighbors model in the validation procedure, what is the RMSE on the test set?

28. **Pre-Owned Car Prices: Neural Networks.** Palanquin Auto is a national pre-owned car retailer that offers an online marketplace for customers interested in purchasing a pre-owned automobile. In an effort to reduce price negotiation, Palanquin collects data on automobile sales in an effort to set fair, competitive prices. The file *palanquin* contains data for the variables listed in Problem 25. Predict the pre-owned car prices using a neural network. Set aside 50% of the data as a test set and use 50% of the data for training and validation. **LO 1, 2, 6**

DATA *file* palanquin

a. Determine the neural network configuration that minimizes the RMSE in a validation procedure.
b. For the best-performing neural network identified in part (a), what is the RMSE on the test set?

29. **Pre-Owned Car Prices: Regularized Linear Regression.** Palanquin Auto is a national pre-owned car retailer that offers an online marketplace for customers interested in purchasing a pre-owned automobile. In an effort to reduce price negotiation, Palanquin collects data on automobile sales in an effort to set fair, competitive prices. The file *palanquin* contains data for the variables listed in Problem 25. Predict the pre-owned car prices using a linear regression model with lasso regularization. Set aside 50% of the data as a test set and use 50% of the data for training and validation. **LO 1, 2, 8**

DATA *file* palanquin

a. Determine the value of the lasso regularization penalty that minimizes the RMSE in a validation procedure.
b. For the best-performing linear regression model identified in part (a), what is the RMSE on the test set?
c. What are the coefficient values for the variables? Does the lasso regularization remove any variables from the model?

Case Problem: Housing Bubble

As an intern with the local home builder's association, you have been asked to analyze the state of the local housing market, which has suffered during a recent economic crisis. You have been provided three data sets: the file *prebubble* which contains information on homes sold during the one-year period before the burst of the "housing bubble," the file

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

postbubble which contains information on homes sold during the one-year period after the burst of the housing bubble, and a file *onmarket* corresponding to homes currently for sale. The variables in the respective data sets are listed in the following table (the file *onmarket* has no Price column as these homes have not sold).

DATA *file*
prebubble.csv
postbubble.csv
onmarket.csv

Variable	Description
LandValue	assessed value for the land ($)
BuildingValue	assessed value for the building structure ($)
Acres	size of lot home sits on (acres)
AboveSpace	above ground living space (sq ft)
Basement	below ground finished living space (sq ft)
Deck	total deck space (sq ft)
Baths	number of full- or 3/4-baths
Toilets	number of 1/2-baths
Fireplaces	number of fireplaces
Beds	number of bedrooms
Rooms	number of rooms, which are not bedrooms
AC	1 if home has air conditioning for at least 1/2 of living space, 0 if not
Age	age of home at time of sale
Car	total space for parking cars in covered structures (attached garage| carport) (sq ft)
Condition	Condition of home (good, average, poor)
Price	price the home was sold for (seasonally adjusted $)

Prepare a managerial report by executing the following steps. **LO 1, 2, 3, 4, 5, 6, 7, 8**

Managerial Report

1. Using the file *prebubble*, train and validate data mining regression models to predict the value of the Price variable.
2. Based on the performance in the validation experiment, select a final predictive model to construct on all the observations in the file *prebubble*. Use this model to predict the prices of the 2,000 homes in the file *onmarket*.
3. Using the file *postbubble*, train and validate data mining regression models to predict the value of the Price variable.
4. Based on the performance in the validation experiment, select a final predictive model to construct on all the observations in the file *postbubble*. Use this model to predict the prices of the 2,000 homes in the file *onmarket*.
5. For each of the 2,000 houses in the test set, compare the predictions from part 2 based on the prebubble data to the predictions from part 4 on the postbubble data. Specifically, compute the percentage difference in predicted price between the prebubble model and postbubble model, where percentage difference = (postbubble predicted price − prebubble predicted price)/(prebubble predicted price). What is the average percentage change in predicted price between the prebubble and postbubble models? What does this suggest about the impact of the housing bubble and its subsequent bursting?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 11

Predictive Data Mining: Classification Tasks

Contents

Analytics in Action: Microsoft

11.1 Data Sampling, Preparation, and Partitioning
- Static Holdout Method
- *k*-Fold Cross-Validation
- Class Imbalanced Data

11.2 Performance Measures for Binary Classification

11.3 Classification with Logistic Regression

11.4 *k*-Nearest Neighbors Classification

11.5 Classification Trees
- Constructing a Classification Tree
- Generating Predictions with a Classification Tree
- Ensemble Methods

11.6 Neural Network Classification
- Structure of a Neural Network
- How a Neural Network Learns

11.7 Feature Selection
- Wrapper Methods
- Filter Methods
- Embedded Methods

Summary 612
Glossary 612
Problems 615

Available in the Cengage eBook:
- Appendix: Classification via Logistic Regression with R
- Appendix: *k*-Nearest Neighbors Classification with R
- Appendix: Individual Classification Trees with R
- Appendix: Random Forests of Classification Trees with R
- Appendix: Neural Network Classification with R
- Appendix: Classification via Logistic Regression with Orange
- Appendix: *k*-Nearest Neighbors Classification with Orange
- Appendix: Individual Classification Trees with Orange
- Appendix: Random Forests of Classification Trees with Orange
- Appendix: Neural Network Classification with Orange

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Evaluate the out-of-sample predictive performance of a model with a static holdout and *k*-fold cross-validation experiment.

LO 2 Compute and interpret performance measures for classification methods.

LO 3 Describe and apply logistic regression to classify observations with binary outcome variables.

LO 4 Describe and apply *k*-nearest neighbors to classify observations with binary outcome variables.

LO 5 Describe and apply classification trees to classify observations with binary outcome variables.

LO 6 Describe and apply ensemble methods to classify observations with binary outcome variables.

LO 7 Describe and apply neural networks to classify observations with binary outcome variables.

LO 8 Describe various ways of selecting features for classification models.

LO 9 Describe the use of regularization to avoid the overfitting of a classification method.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

Microsoft*

While the widespread use of the internet to conduct business provides many benefits to merchants, it also exposes them to fraudulent activity such as e-commerce payment fraud, free-trial abuse, fake product reviews, warranty fraud, and refund fraud. Accounting for the cost of managing fraud events, it is estimated that e-commerce merchants lose over $250 billion annually due to fraud. E-commerce payment fraud is of particular concern to vendors such as Microsoft. An example of e-commerce payment fraud is the use of a stolen credit card number to purchase products or services. A common application of classification models is to try to use the activity pattern of a customer to determine whether it corresponds to fraudulent behavior.

To protect e-commerce business, Microsoft has developed an effective fraud-management system. One of the challenges facing Microsoft's fraud detection team is that the activity patterns associated with fraudulent behavior change over time as a consequence of the fraudsters varying their tactics. In addition, legitimate customers' online shopping behaviors also change over time. To address this, the Microsoft fraud detection team used both a long-term predictive model and short-term predictive model. The long-term model captures the general trend of fraud and is trained weekly on at least one year's worth of data. The short-term model then uses the output of the long-term model as one of its input features and is designed to react quickly to newly emerging fraud patterns. The short-term model is trained on the most recent data with both confirmed and unconfirmed fraud cases. The final result is an estimate of the probability that a given activity pattern corresponds to fraud. By improving fraud detection, Microsoft has lowered the rate of successful e-commerce payment fraud and improved the customer experience.

Nanduri, J, Jia, Y., Oka, A., Beaver, J. and Y. Liu, "Microsoft Uses Machine Learning and Optimization to Reduce E-Commerce Fraud," *INFORMS Journal on Applied Analytics*, vol. 50, no. 1 (2020): 64–79.

In this chapter, we focus on data mining methods for **classification**, the task of predicting categorical outcome variable using a set of input variables, or **features**. Specifically, we focus on classification tasks in which the categorical outcome variable has two possible categories. These are called binary classification tasks, and they apply to a wide range of situations.

In Chapter 10, we discuss data mining methods for addressing the task of predicting an outcome variable when it is quantitative instead of categorical.

Predictive data mining methods for classification tasks are a type of **supervised learning** because the known values of outcome variable "supervise" or guide the process of learning how to use the input features to predict future values of the outcome variable.

The data mining process comprises the following steps:

1. *Data sampling*. Extract a sample of data that is relevant to the business problem under consideration.
2. *Data preparation*. Manipulate the data to put it in a form suitable for formal modeling. This step includes addressing missing and erroneous data, reducing the number of variables, and defining new variables. Data exploration is an important part of this step and may involve the use of descriptive statistics, data visualization, and clustering to better understand the relationships supported by the data.
3. *Data partitioning*. Divide the sample data for the training, validation, and testing of the data mining algorithm performance.
4. *Model Construction and Assessment.* Construct a set of supervised learning models (e.g., logistic regression, *k*-nearest neighbors, tree-based methods, neural networks) on the training data and make out-of-sample predictions on the validation set to compare their performance. Select a final model and make predictions on the test set as a final appraisal of the final model's performance.

In Chapters 2 and 3, we discuss descriptive statistics and data-visualization techniques. The data preparation process is discussed in Chapter 4 and dimension reduction and clustering techniques are discussed in Chapter 6.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

11.1 Data Sampling, Preparation, and Partitioning

Upon identifying a business problem, data on relevant variables must be obtained for analysis. Data repositories with millions of observations over hundreds of measured variables are now common. If the volume of relevant data is extremely large (thousands of observations or more), it is unnecessary (and computationally challenging) to use all the data in order to perform a detailed analysis. When dealing with large volumes of data (with hundreds of thousands or more observations), best practice is to extract a representative sample (with thousands or tens of thousands of observations) for analysis. A sample is representative if the analyst can draw the same conclusions from the sample as the analyst would draw from the entire population of data.

There are no definite rules to determine the size of a sample. The sample of data must be large enough to contain substantial information, yet small enough to process and analyze quickly. If the sample is too small, relationships in the data may be missed or spurious relationships may be suggested. Perhaps the best advice is to use enough data to eliminate any doubt about whether the sample size is sufficient because data mining algorithms typically are more effective given more data.

In Chapter 10, we present an example based on a regression task that motivates data partitioning for predictive models via the static holdout method and k-fold cross-validation.

In predictive modeling, the goal is to construct a model that predicts future observations of the currently unknown outcome variable as accurately as possible. To facilitate this, once a representative data sample has been prepared for analysis, it must be partitioned to appropriately evaluate the performance of one or more predictive models. In this chapter, we construct classification models on one set of observations and then assess their performance on a separate set of observations. This experimental approach helps to guard against the selection of overfit models. For a binary classification task, **overfitting** occurs when the model does a good job predicting when an event does and does not occur for data on which the model was constructed, but does a worse job predicting when an event does and does not occur for data not included in the model construction process. In the rest of this section, we describe two common methods for partitioning the data for model construction and assessment, as well as discussing the situation in which one of the outcomes for a binary outcome variable occurs much less frequently than the other.

Static Holdout Method

This is called a static holdout method because the partitioning of the data is done only once. In contrast, as we describe in the next section, the k-fold cross-validation procedure creates k different partitions that are iteratively used for training and validation.

The most straightforward way to partition data to combat overfitting is to split the sample data into three static data sets: the training set, the validation set, and the test set. The **training set** consists of the data used to build the candidate classification models. However, since the training set consists of the data used to build the models, it cannot be used to reliably identify the best model for data outside of the training set. Therefore, we use classification models constructed on the training set to make predictions on the **validation set**, which is a set of observations that have not been used in model training. Candidate classification models are then compared based on their predictive performance on the validation set.

The validation set is typically used to select a final model through either comparison with other models or the tuning of model parameters. As a result of choosing the classification model with the "best" performance on the validation set, this final model's performance measures on the validation set tend to be optimistic estimates of its future performance. Thus, improved estimates of the final model's future performance may be obtained by applying the model to a third set of data, called the **test set**.

For example, suppose we have constructed four models on training data and compared their performance on validation data. After identifying the best of the four models based on their performances on validation data, we apply this final model to a test set (consisting of observations that were not part of the training or validation data) in order to obtain an unbiased estimate of this model's performance on future applications. The final model's performance on the test set helps us understand how well we can expect this model to perform in the future.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

There are no definite rules for the size of the three partitions, but the training set is typically the largest. For a binary classification task, a rule of thumb is to have at least $12 \times q$ observations, where q is the number of variables.

The advantages of the static holdout method are that it is simple and quick. However, the results of the modeling procedure based on a single application of the static holdout can vary greatly depending on which observations and how many observations are selected for the training and validation sets. If we would construct another iteration of the static holdout method and split the observations differently across the sets, we may find that a different model performs best. Motivated by this realization, we describe a more robust data partitioning method to train and validate models in the next section.

k-Fold Cross-Validation

In the static holdout method, the observations that are placed into the training and validation sets can have an impact on how models are built and evaluated. In this section, we discuss a more robust procedure to train and validate models called **k-fold cross-validation**. As a pre-processing step, if an unbiased estimate of the final model is desired, we first remove a desired number of observations to compose the test set. The remaining observations will be used to train and validate the model in an iterative procedure. The k-fold cross-validation procedure randomly divides the data not in the test set into k equal-sized (or nearly equal-sized), mutually exclusive, and collectively exhaustive subsets of observations called folds. The cross-validation procedure consists of k iterations in which a different subset is designated as the validation set at each iteration and the remaining $k - 1$ subsets are combined and designated as the training set. A common choice for the number of folds is $k = 10$. As Figure 11.1 shows, a 10-fold cross-validation consists of 10 iterations. At each iteration of a 10-fold cross-validation, a different collection of 9 folds (90% of the data) is used to construct the predictive model which is then evaluated on the remaining unused fold (10% of the data). At the end of the $k = 10$ iterations, each observation in the data set has been predicted exactly once by a version of the predictive model trained on the 9 folds not containing the observation. This results in an estimate of the out-of-sample performance of the predictive model that can be used to select the final model. The k-fold cross-validation method is more complex and time consuming than the static holdout method, but the results of the k-fold cross-validation method are less sensitive to which observations are used in the training and the validation sets, respectively.

In the remainder of this chapter, our examples demonstrating different predictive models for a regression task will use the static holdout method for simplicity. However, k-fold cross-validation is recommended in practice.

Leave-one-out cross-validation is a special case of k-fold cross validation in which each fold consists of a single observation. That is, for training and validation on a set of n observations, an iteration of leave-one-out cross-validation consists of estimating the model on $n - 1$ observations and evaluating the model on the single observation that was omitted from the training data. This procedure is repeated for n total iterations so that the model is trained on each possible combination of $n - 1$ observations and evaluated on the single remaining observation in each case.

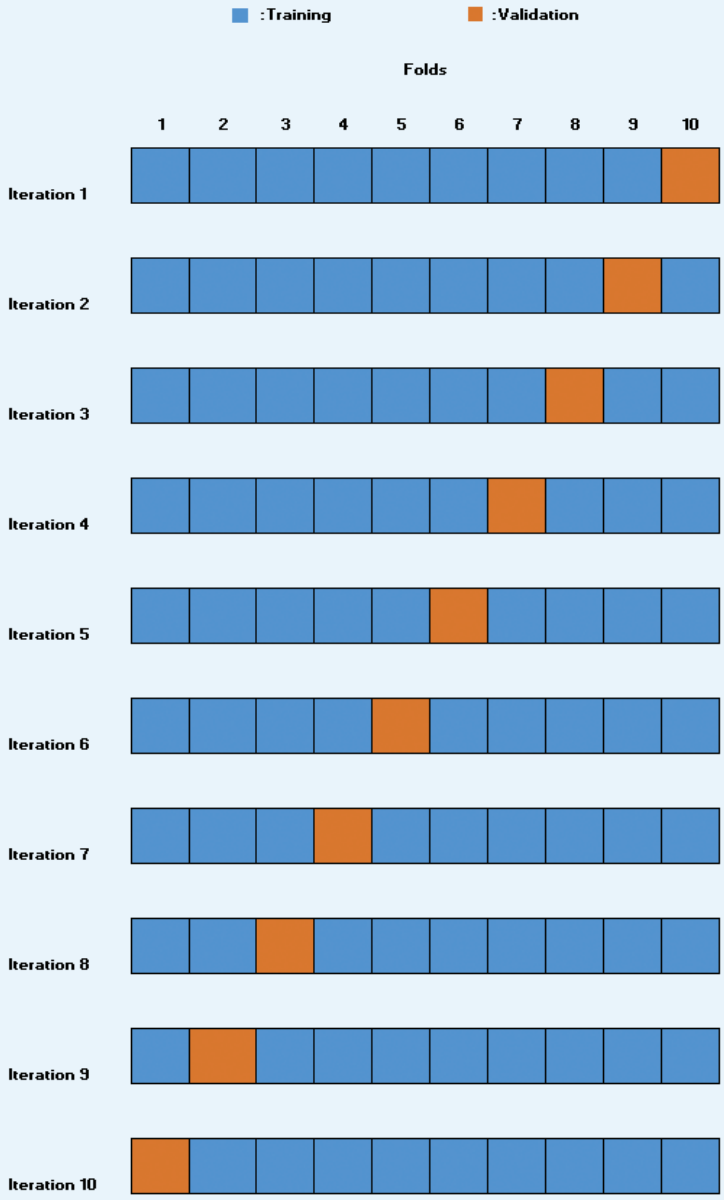
Class Imbalanced Data

The classification of rare events (e.g., click-through on an advertisement posted on a web site, claims made against property-insurance policies, or fraudulent credit card charges) presents a number of challenges. In a binary classification problem, the identification of the rare class often has greater value than identifying the frequent class. As we will discuss in the following section on performance measures, there are a variety of metrics that can be used to evaluate a classifier's effectiveness in correctly identifying observations from the rare class.

Another approach to handling a severe imbalance between the two observation classes is to modify the sampling mechanism used to construct the training data. For example, suppose we are trying to construct a classifier to identify whether or not a credit card charge is fraudulent. If 0.1% of all credit card charges are fraudulent, then a random sample of

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.1 Demonstration of a 10-Fold Cross-Validation Experiment



500,000 observations would contain only approximately 500 observations corresponding to fraudulent charges. In a case such as this, it is common to create training, validation, and test sets through stratified random sampling. For simplicity, consider a static holdout approach splitting 64% of the data into a training set, 16% of the data into a validation set, and 20% of the data into a test set. **Stratified random sampling** will ensure that 0.1% of each of these three sets will consist of fraudulent charge observations. That is, (1) the training set will contain 320,000 observations, 320 of which correspond to fraudulent charges, (2) the validation set will containing 80,000 observations, 80 of which correspond to fraudulent charges, and (3) the test set will contain 100,000 observations, 100 of which correspond to fraudulent charges.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notice that when using stratified sampling to construct the training set in the credit card fraud example, only 320 observations correspond to fraudulent charges while 319,680 correspond to non-fraudulent charges. In an attempt to balance the class information used to train a classification model, there are two basic sampling approaches for modifying the class distribution of the training set: undersampling and oversampling. In each of these approaches, the validation set and test set are constructed to be representative of the original data via stratified sampling, but the percentage of minority class observations in the training set is increased.

An **undersampling** approach reduces the size of the training set by considering fewer majority class (non-fraudulent) observations in order to balance the class distribution. For the credit card fraud example, undersampling could obtain a balanced training set of 640 observations consisting of 320 fraudulent observations and 320 non-fraudulent observations.

In contrast, an **oversampling** approach maintains a larger training set size by sampling the minority class observations with replacement—effectively making copies of the minority class observations. For the credit card fraud example, a balanced training set of 2,000 observations can be achieved by considering 1,000 randomly selected fraudulent observations (many of which are duplicates) and 1,000 randomly selected non-fraudulent observations.

Both undersampling and oversampling have drawbacks. By removing the majority class observations, a training set constructed via undersampling is susceptible to missing important information about the majority class. This can be addressed by performing undersampling multiple times to create multiple classifiers. By inserting copies of minority class observations, a training set constructed via oversampling may contain a lot of duplicated noise (rather than information) which may cause model overfitting. This can be addressed by careful use of a validation set to evaluate classification models.

Undersampling is often utilized when there is a lot of data from which to sample (e.g., more than 10,000 observations), while oversampling is often utilized when there is less data from which to sample (e.g., less than 10,000 observations). However, the choice of undersampling or oversampling is also affected by the degree of class imbalance.

Notes + Comments

There are additional sampling-based approaches to modify the class distribution of the training set when considering a data set with class imbalance. There are hybrid approaches which combine undersampling and oversampling as well as more sophisticated approaches. For example, an approach referred to as SMOTE (synthetic minority over-sampling technique) oversamples the minority class, but instead of simply copying minority class observations, it creates variation in these copied observations by randomly perturbing one or more of their attribute values.

11.2 Performance Measures for Binary Classification

DATA*file* optiva

There are several different ways to measure the performance of a classification model. We describe each of these in the context of an example from the financial services industry. Optiva Credit Union wants to better understand its personal lending process and its loan customers. The file *optiva* contains over 40,000 customer observations with information on whether the customer defaulted on a loan, customer age, average checking account balance, whether the customer had a mortgage, the customer's job status, the customer's marital status, and the customer's level of education. We will use these data to demonstrate how to measure the ability of a classification method to identify customers who default on their loans.

In our treatment of classification problems, we restrict our attention to problems for which we want to classify observations into one of two possible classes (e.g., loan default or no default), but the concepts generally extend to cases with more than two classes. A natural way

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

to evaluate the performance of a classification method, or classifier, is to count the number of times that an observation is predicted to be in the wrong class. By counting the classification errors on a sufficiently large validation set and/or test set that is representative of the population, we will generate an accurate measure of classification performance of our model.

Classification error is commonly displayed in a **confusion matrix**, which displays a model's correct and incorrect classifications. Table 11.1 illustrates a confusion matrix resulting from an attempt to classify the customer observations in a subset of data from the file *optiva*. In this table, Class 1 = loan default and Class 0 = no default. The confusion matrix is a cross-tabulation of the actual class of each observation and the predicted class of each observation. From the first row of the matrix in Table 11.1, we see that 7,479 observations corresponding to nondefaults were correctly identified and 5,244 actual nondefault observations were incorrectly classified as loan defaults. From the second row of Table 11.1, we observe that 89 actual loan defaults were incorrectly classified as nondefaults and 146 observations corresponding to loan defaults were correctly identified.

Many measures of classification performance are based on the confusion matrix. The percentage of misclassified observations is expressed as the **overall error rate** and is computed as

$$\text{Overall error rate} = \frac{n_{10} + n_{01}}{n_{11} + n_{10} + n_{01} + n_{00}}$$

where n_{10}, n_{01}, n_{11}, and n_{00} are as defined in Table 11.1. The overall error rate of the classification in Table 11.1 is $(89 + 5{,}244)/(146 + 89 + 5{,}244 + 7{,}479) = 41.2\%$. One minus the overall error rate is often referred to as the **accuracy** of the model. The model accuracy based on Table 11.1 is 58.8%.

While overall error rate conveys an aggregate measure of misclassification, it counts misclassifying an actual Class 0 observation as a Class 1 observation (a **false positive**) the same as misclassifying an actual Class 1 observation as a Class 0 observation (a **false negative**). In many situations, the cost of making these two types of errors is not equivalent. For example, suppose we are classifying patient observations into two categories: Class 1 is cancer and Class 0 is healthy. The cost of incorrectly classifying a healthy patient observation as "cancer" will likely be limited to the expense and stress of additional testing. The cost of incorrectly classifying a cancer patient observation as "healthy" may result in an indefinite delay in treatment of the cancer and premature death of the patient.

To account for the asymmetric costs in misclassification, we define the error rate with respect to the individual classes:

$$\textbf{Class 1 error rate} = \frac{n_{10}}{n_{11} + n_{10}}$$

$$\textbf{Class 0 error rate} = \frac{n_{01}}{n_{01} + n_{00}}$$

The Class 1 error rate of the classification in Table 11.1 is $89/(146 + 89) = 37.9\%$. The Class 0 error rate of the classification in Table 11.1 is $(5{,}244)/(5{,}244 + 7{,}479) = 41.2\%$. That is, the model that produced the classifications in Table 11.1 is slightly better at predicting Class 1 observations than Class 0 observations.

In Table 11.1, n_{01} is the number of false positives and n_{10} is the number of false negatives.

Table 11.1 Confusion Matrix

	Predicted Class	
Actual Class	**0**	**1**
0	$n_{00} = 7{,}479$	$n_{01} = 5{,}244$
1	$n_{10} = 89$	$n_{11} = 146$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To understand the tradeoff between Class 1 error rate and Class 0 error rate, we must be aware of the criteria generally used by classification algorithms to classify observations. Most classification algorithms first estimate an observation's probability of Class 1 membership and then classify the observation into Class 1 if this probability meets or exceeds a specified **cutoff value**. Typically, the default cutoff value for a classifier is 0.5. Modifying the cutoff value affects the likelihoods of occurrence for these two types of classification error. As we decrease the cutoff value, more observations will be classified as Class 1, thereby increasing the likelihood that a Class 1 observation will be correctly classified as Class 1; that is, Class 1 error will decrease. However, as a side effect, more Class 0 observations will be incorrectly classified as Class 1; that is, Class 0 error will rise.

To demonstrate how the choice of cutoff value affects classification error, Table 11.2 shows a list of 50 observations (11 of which are actual Class 1 members) and an estimated probability of Class 1 membership produced by the classification algorithm. Table 11.3 shows the confusion matrices and corresponding Class 1 error rates, Class 0 error rates, and overall error rates for cutoff values of 0.75, 0.5, and 0.25, respectively. As we decrease the cutoff value, more observations will be classified as Class 1, thereby increasing the likelihood that a Class 1 observation will be correctly classified as Class 1 (decreasing the Class 1 error rate). However, as a side effect, more Class 0 observations will be incorrectly classified as Class 1 (increasing the Class 0 error rate). That is, we can accurately identify more of the actual Class 1 observations by lowering the cutoff value, but we do so at a cost of

Table 11.2 Classification Probabilities

Actual Class	Probability of Class 1	Actual Class	Probability of Class 1
1	1.00	0	0.66
1	1.00	0	0.65
0	1.00	1	0.64
1	1.00	0	0.62
0	1.00	0	0.60
0	0.90	0	0.51
1	0.90	0	0.49
0	0.88	0	0.49
0	0.88	1	0.46
1	0.88	0	0.46
0	0.87	1	0.45
0	0.87	1	0.45
0	0.87	0	0.45
0	0.86	0	0.44
1	0.86	0	0.44
0	0.86	0	0.30
0	0.86	0	0.28
0	0.85	0	0.26
0	0.84	1	0.24
0	0.84	0	0.22
0	0.83	0	0.21
0	0.68	0	0.04
0	0.67	0	0.04
0	0.67	0	0.01
0	0.67	0	0.00

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 11.3 Confusion Matrices for Various Cutoff Values

Cutoff Value = 0.75

	Predicted Class	
Actual Class	**0**	**1**
0	$n_{00} = 24$	$n_{01} = 15$
1	$n_{10} = 5$	$n_{11} = 6$

Actual Class	No. of Cases	No. of Errors	Error Rate (%)
0	$n_{00} + n_{01} = 39$	$n_{01} = 15$	38.46
1	$n_{10} + n_{11} = 11$	$n_{10} = 5$	45.45
Overall	$n_{00} + n_{01} + n_{10} + n_{11} = 50$	$n_{01} + n_{10} = 20$	40.00

Cutoff Value = 0.50

	Predicted Class	
Actual Class	**0**	**1**
0	$n_{00} = 15$	$n_{01} = 24$
1	$n_{10} = 4$	$n_{11} = 7$

Actual Class	No. of Cases	No. of Errors	Error Rate (%)
0	39	24	61.54
1	11	4	36.36
Overall	50	28	56.00

Cutoff Value = 0.25

	Predicted Class	
Actual Class	**0**	**1**
0	$n_{00} = 6$	$n_{01} = 33$
1	$n_{10} = 1$	$n_{11} = 10$

Actual Class	No. of Cases	No. of Errors	Error Rate (%)
0	39	33	84.62
1	11	1	9.09
Overall	50	34	68.00

misclassifying more actual Class 0 observations as Class 1 observations. Figure 11.2 shows the Class 1 and Class 0 error rates for cutoff values ranging from 0 to 1. One common approach to handling the tradeoff between Class 1 and Class 0 error is to set the cutoff value to minimize the Class 1 error rate subject to a threshold on the maximum Class 0 error rate. Specifically, Figure 11.2 illustrates that for a maximum allowed Class 0 error rate of 70%, a cutoff value of 0.45 (depicted by the vertical dashed line) achieves a Class 1 error rate of 20%.

As we have mentioned, identifying Class 1 members is often more important than identifying Class 0 members. One way to evaluate a classifier's value is to compare its effectiveness in identifying Class 1 observations as compared with random classification. To gauge a classifier's added value, a **lift chart** compares the number of actual Class 1 observations identified if considered in decreasing order of their estimated probability of being

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.2 Classification Error Rates vs. Cutoff Value

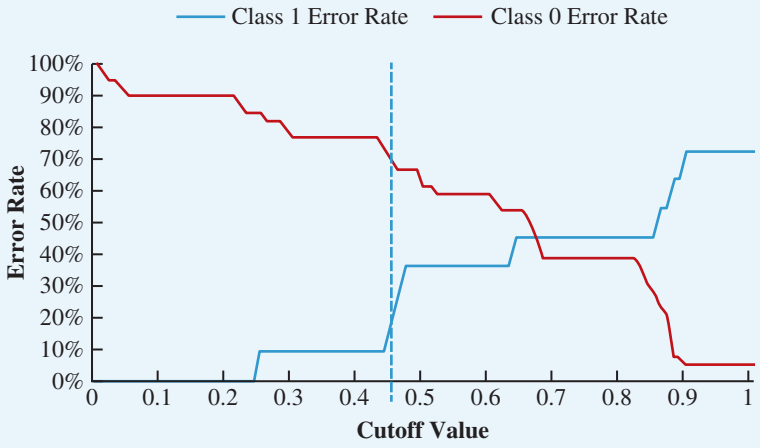


Figure 11.2 was created using a Data Table that varied the cutoff value and tracked the Class 1 error rate and Class 0 error rate. For instructions on how to construct Data Tables in Excel, see Chapter 12.

in Class 1 to the number of actual Class 1 observations identified if randomly selected. Formally, the lift corresponding to the n observations with the largest estimated probability of being in Class 1 is computed as

$$\text{lift}(n) = \frac{\text{actual number of Class 1 observations in top } n \text{ observations with largest } P(\text{Class 1})}{\text{expected number of Class 1 observations in } n \text{ randomly selected observations}}$$

Referring to Table 11.2, we observe that 11 of the 50 observations are Class 1 observations. Therefore, the expected number of Class 1 observations in n randomly selected observations is $(11/50) \times n$. Let's consider the lift corresponding to $n = 10$ observations with the largest estimated probability of being in Class 1. From Table 11.2, we see that in the ten observations with the largest estimated probability of being in Class 1, there are five actual Class 1 observations. Thus, for this example,

$$\text{lift}(10) = \frac{5}{\frac{11}{50} \times 10} = 2.273$$

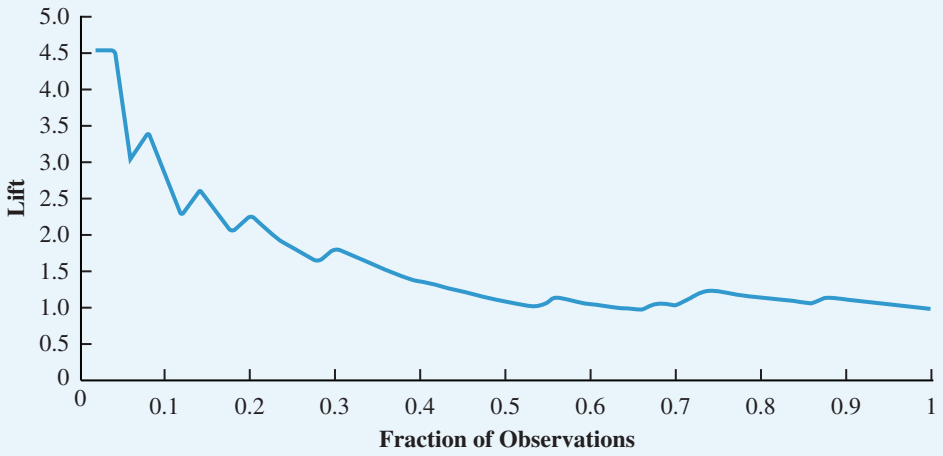
In its ten most likely observations to be Class 1, the classifier identifies 2.273 times more actual Class 1 observations than randomly selecting ten observations. The lift curve in Figure 11.3 plots the values of lift(n) for values of n ranging from 1 to the total number of observations, where the horizontal axis expresses the number of observations being considered as a fraction of the total number observations. For example, lift(10) = 2.273 is plotted as the point (0.2, 2.273) because the fraction of the observations being considered is 10/50 = 0.2. A lift curve generally trends downward as a larger fraction of observations is considered because as one considers a larger fraction of the observations, the benefit of using a classifier versus random guessing decreases. At the extreme, if all the observations are considered, then the lift of the classifier is 1.0 corresponding to the point (1.0, 1.0) in Figure 11.3.

Lift charts are prominently used in direct-marketing applications that seek to identify customers who are likely to respond to a targeted promotion. In these applications, it is common to have a fixed budget and, therefore, a fixed number of customers to target. Lift charts identify how much better a data mining model does at identifying responders than targeting a random set of customers.

In addition to the overall error rate, Class 1 error rate, and Class 0 error rate, there are other measures that gauge a classifier's performance. The ability to correctly predict

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.3 Lift Chart



Class 1 (positive) observations is expressed by subtracting the Class 1 error rate from 1. The resulting measure is referred to as the **sensitivity**, or **recall**, which is calculated as

$$\text{Sensitivity} = 1 - \text{Class 1 error rate} = \frac{n_{11}}{n_{11} + n_{10}}$$

Similarly, the ability to correctly predict Class 0 (negative) observations is expressed by subtracting the Class 0 error rate from 1. The resulting measure is referred to as the **specificity**, which is calculated as

$$\text{Specificity} = 1 - \text{Class 0 error rate} = \frac{n_{00}}{n_{00} + n_{01}}$$

The sensitivity of the model that produced the classifications in Table 11.1 is $146/(146 + 89) = 62.1\%$. The specificity of the model that produced the classifications in Table 11.1 is $7{,}479/(5{,}244 + 7{,}479) = 58.8\%$.

Precision is a measure that corresponds to the proportion of observations predicted to be Class 1 by a classifier that are actually in Class 1

$$\text{Precision} = \frac{n_{11}}{n_{11} + n_{01}}$$

The **F1 Score** combines precision and sensitivity into a single measure and is defined as

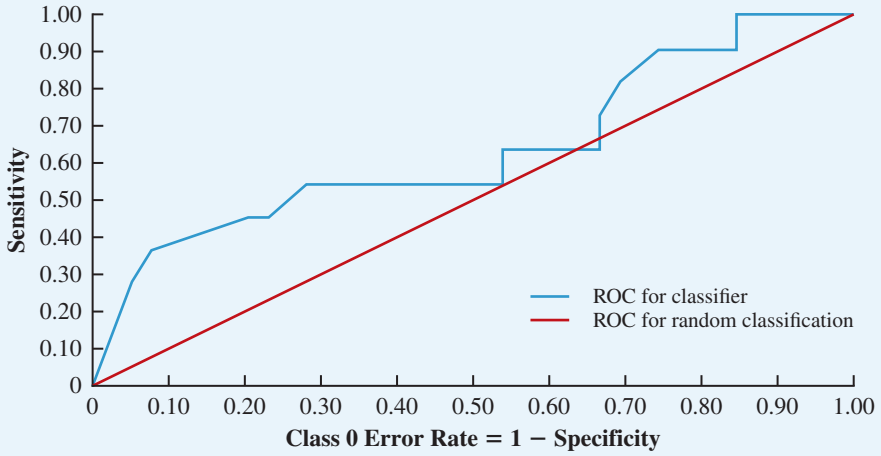
$$\text{F1 score} = \frac{2n_{11}}{2n_{11} + n_{01} + n_{10}}$$

As we illustrated in Figure 11.2, decreasing the cutoff value will decrease the number of actual Class 1 observations misclassified as Class 0, but at the cost of increasing the number of Class 0 observations that are misclassified as Class 1. The **receiver operating characteristic (ROC) curve** is an alternative graphical approach for displaying this tradeoff between a classifier's ability to correctly identify Class 1 observations and its Class 0 error rate. In a ROC curve, the vertical axis is the sensitivity of the classifier, and the horizontal axis is the Class 0 error rate (which is equal to 1 − specificity).

In Figure 11.4, the blue curve depicts the ROC curve corresponding to the classification probabilities for the 50 observations in Table 11.2. The red diagonal line in Figure 11.4 represents the expected sensitivity and Class 0 error rate achieved by random classification of the 50 observations. The point (0, 0) on the blue curve occurs when the cutoff value is set so that

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.4 Receiver Operating Characteristic (ROC) Curve



all observations are classified as Class 0; for this set of 50 observations, a cutoff value greater than 1.0 will achieve this. That is, for a cutoff value greater than 1, for the observations in Table 11.2, sensitivity $= 0/(0 + 11) = 0$ and the Class 0 error rate $= 0/(0 + 39) = 0$. The point (1, 1) on the curve occurs when the cutoff value is set so that all observations are classified as Class 1; for this set of 50 observations, a cutoff value of zero will achieve this. That is, for a cutoff value of 0, sensitivity $= 11/(11 + 0) = 1$ and the Class 0 error rate $= 39/(39 + 0) = 1$. Repeating these calculations for varying cutoff values and recording the resulting sensitivity and Class 0 error rate values, we can construct the ROC curve in Figure 11.4.

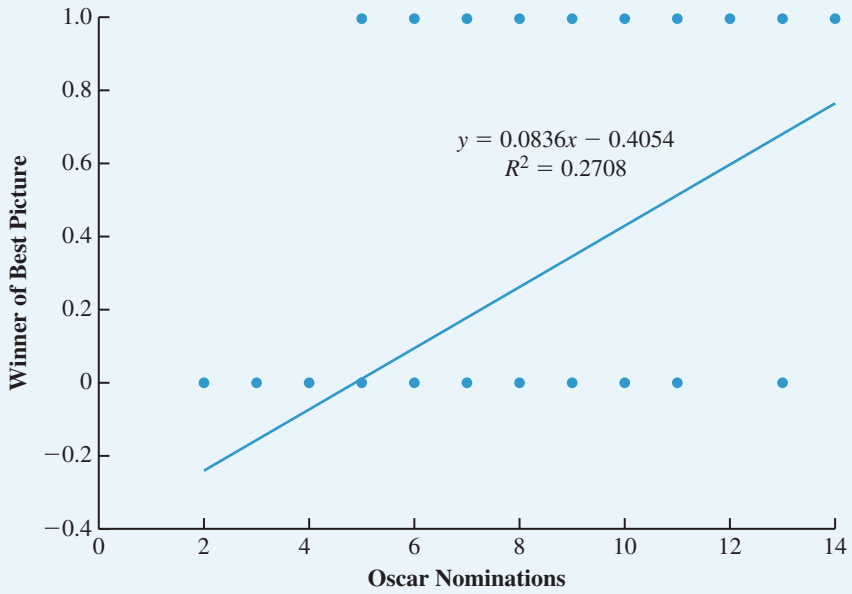
In general, we can evaluate the quality of a classifier by computing the **area under the ROC curve**, often referred to as the AUC. The greater the area under the ROC curve, i.e., the larger the AUC, the better the classifier performs. To understand why, suppose there exists a cutoff value such that a classifier correctly identifies each observation's actual class. Then, the ROC curve will pass through the point (0, 1), which represents the case in which the Class 0 error rate is zero and the sensitivity is equal to one (which means that the Class 1 error rate is zero). In this case, the area under the ROC curve would be equal to one as the curve would extend from (0, 0) to (0, 1) to (1, 1). In Figure 11.4, note that the area under the red diagonal line representing random classification results is 0.5. In Figure 11.4, we observe that the classifier is providing value over a random classification, as its AUC is greater than 0.5. The closer a classifier's AUC is to 1.0, the better the classifier is at separating Class 1 observation from Class 0 observations. The closer a classifier's AUC is to 0.5, the more similar the classifier is to a random classification.

11.3 Classification with Logistic Regression

Similar to how multiple linear regression predicts a continuous outcome variable, y, with a collection of explanatory variables, $x_1, x_2, \ldots, x_q$, via the linear equation $\hat{y} = b_0 + b_1x_1 + \cdots + b_qx_q$, **logistic regression** attempts to classify a binary categorical outcome ($y = 0$ or 1) as a linear function of explanatory variables. However, directly trying to explain a binary outcome via a linear function of the explanatory variables is not effective. To understand this, consider the task of predicting whether a movie wins the Academy Award for Best Picture using information on the total number of Oscar nominations that a movie has received. Figure 11.5 shows a scatter chart of a sample

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.5 Scatter Chart and Simple Linear Regression Fit for Oscars Example



DATA*file* academy

of movie data found in the file *academy*; each data point corresponds to the total number of Oscar nominations that a movie received and whether the movie won the best picture award (1 = movie won, 0 = movie lost). The line in Figure 11.5 corresponds to the simple linear regression fit. This linear function can be thought of as predicting the probability p of a movie winning the Academy Award for Best Picture via the equation $\hat{p} = -0.4054 + 0.836 \times$ Oscar Nominations. As Figure 11.5 shows, a linear regression model fails to appropriately explain a binary outcome variable. This model predicts that a movie with fewer than 5 total Oscar nominations has a negative probability of winning the best picture award. For a movie with more than 17 total Oscar nominations, this model predicts a probability greater than 1.0 of winning the best picture award. Furthermore, the residual plot in Figure 11.6 shows an unmistakable pattern of systematic misprediction, suggesting that the simple linear regression model is not appropriate.

As we discuss in Chapter 8, if a linear regression model is appropriate, the residuals should appear randomly dispersed with no discernible pattern.

Estimating the probability p with the linear function $\hat{p} = b_0 + b_1x_1 + \cdots + b_qx_q$ does not fit well because, although p is a continuous measure, it is restricted to the range [0, 1]; that is, a probability cannot be less than zero or larger than one. Figure 11.7 shows an S-shaped curve that appears to better explain the relationship between the probability p of winning the best picture award and the total number of Oscar nominations. Instead of extending off to positive and negative infinity, the S-shaped curve flattens and never goes above one or below zero. We can achieve this S-shaped curve by estimating an appropriate function of the probability p of winning the best picture award with a linear function rather than directly estimating p with a linear function.

As a first step in representing the probability of an event with an S-shaped curve, we note that there is a measure related to probability known as *odds* that is very prominent in gambling and epidemiology. If an estimate of the probability of an event is $\hat{p}$ then the equivalent odds measure is $\hat{p}/(1 - \hat{p})$. For example, if the probability of an event is $\hat{p} = 2/3$, then the odds measure would be $(2/3)/(1/3) = 2$, meaning that the odds are 2 to 1 that the event will occur. The odds metric ranges between zero and positive infinity, so by considering the odds measure rather than the probability $\hat{p}$, we eliminate the linear fit

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.6 Residuals for Simple Linear Regression on Oscars Data

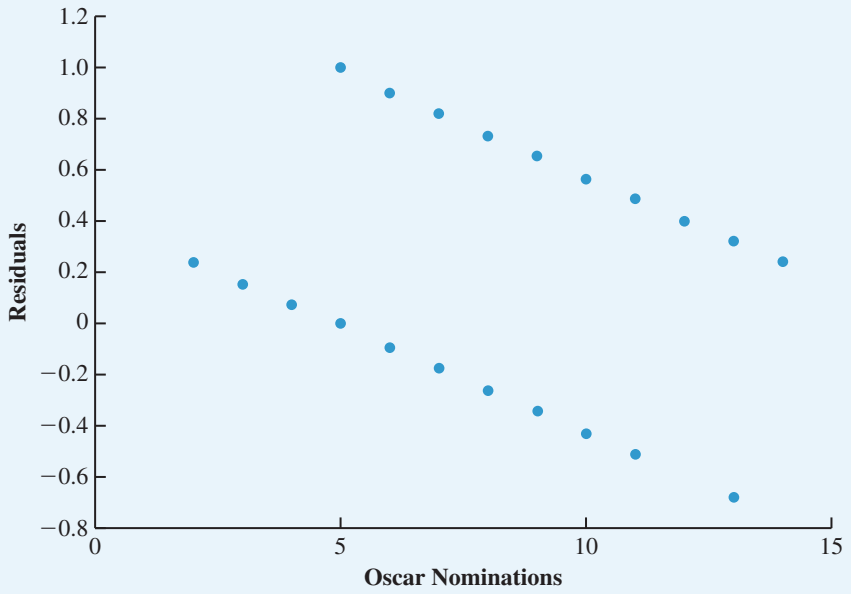
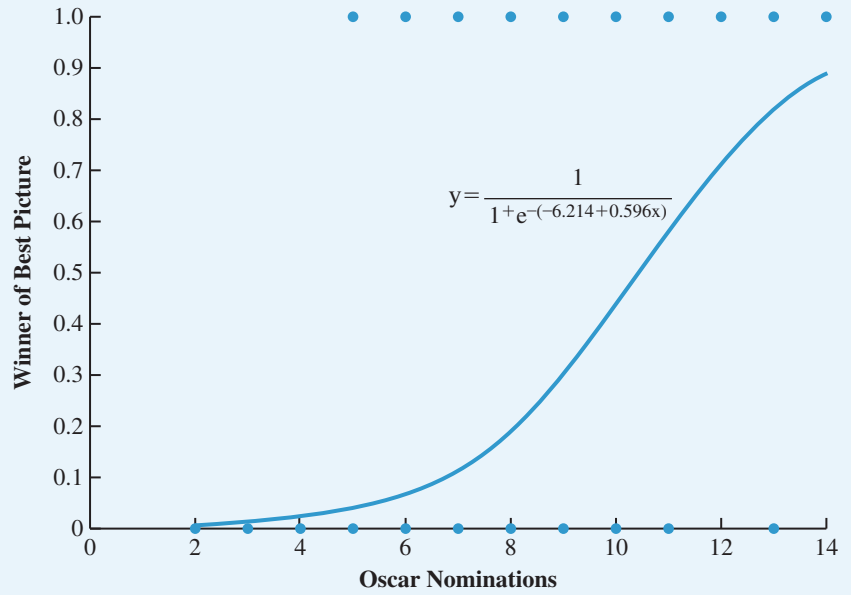


Figure 11.7 Logistic *S*-Curve for Oscars Example



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

problem resulting from the upper bound of one on the probability $\hat{p}$. To eliminate the fit problem resulting from the remaining lower bound of zero on $\hat{p}/(1-\hat{p})$, we observe that the natural log of the odds for an event, also known as "log odds" or logit, $\ln(\hat{p}/(1-\hat{p}))$, ranges from negative infinity to positive infinity. Estimating the logit with a linear function results in a logistic regression model:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_0 + b_1x_1 + \cdots + b_qx_q \tag{11.1}$$

Given a training set of observations consisting of values for a set of explanatory variables, $x_1, x_2, \ldots, x_q$, and whether or not an event of interest occurred ($y = 0$ or 1), the logistic regression model fits values of $b_0, b_1, \ldots, b_q$ that best estimate the log odds of the event occurring. Using statistical software to fit the logistic regression model to the data in the file *academy* results in estimates of $b_0 = -6.214$ and $b_1 = 0.596$; that is, the log odds of a movie winning the best picture award is given by

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -6.214 + 0.596 \text{ Oscar Nominations} \tag{11.2}$$

Unlike the coefficients in a multiple linear regression, the coefficients in a logistic regression do not have an intuitive interpretation. For example, $b_1 = 0.596$ means that for every additional Oscar nomination that a movie receives, its log odds of winning the best picture award increase by 0.596. In other words, the total number of Oscar nominations is linearly related to the log odds of a movie winning the best picture award. Unfortunately, a change in the log odds of an event is not as easy as to interpret as a change in the probability of an event. Algebraically solving equation (11.1) for p, we can express the relationship between the estimated probability of an event and the explanatory variables with an equation known as the logistic function:

Logistic Function

$$\hat{p} = \frac{1}{1 + e^{-(b_0 + b_1x_1 + \cdots + b_qx_q)}} \tag{11.3}$$

Using the data in the file *academy* to estimate the probability of a movie winning the best picture award with its total number of Oscar nominations as the input variable, equation (11.3) is

$$\hat{p} = \frac{1}{1 + e^{-(-6.214 + 0.596 \text{ Oscar Nominations})}} \tag{11.4}$$

Plotting equation (11.4), we obtain the S-shaped curve of Figure 11.7. Clearly, the logistic regression fit implies a nonlinear relationship between the probability of winning the best picture award and the total number of Oscar nominations. The effect of increasing the total number of Oscar nominations on the probability of winning the best picture award depends on the original number of Oscar nominations. For instance, if the total number of Oscar nominations is four, an additional Oscar nomination increases the estimated probability of winning the best picture award from $\hat{p} = \frac{1}{1 + e^{-(-6.214 + 0.596 \times 4)}} = 0.021$ to $\hat{p} = \frac{1}{1 + e^{-(-6.214 + 0.596 \times 5)}} = 0.038$, an increase of 0.017. But if the total number of Oscar nominations is eight, an additional Oscar nomination increases the estimated probability of winning the best picture award from $\hat{p} = \frac{1}{1 + e^{-(-6.214 + 0.596 \times 8)}} = 0.191$ to $\hat{p} = \frac{1}{1 + e^{-(-6.214 + 0.596 \times 9)}} = 0.299$, an increase of 0.108.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

As with other classification methods, logistic regression classifies an observation by using equation (11.3) to compute the probability of an observation belonging to Class 1 and then comparing this probability to a cutoff value. If the probability exceeds the cutoff value (a typical value is 0.5), the observation is classified as Class 1 and otherwise it is classified as Class 0. Table 11.4 shows a subsample of the predicted probabilities computed using equation (11.4) and the subsequent classification for a cutoff value of 0.5.

To demonstrate how a logistic regression model classifies observations when considering two input variables, let us reconsider the Oscars example. In this logistic regression model, we will use the Oscar Nominations and Golden Globe Wins as input variables. The Golden Globe Awards are a motion picture award show held before the Academy Awards and the number of awards that a movie wins at the Golden Globe Awards may be predictive of its success at the Academy Awards. Using data in the file *academy*, the estimated logistic regression model is

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -6.44 + 0.458 \text{ Oscar Nominations} + 0.762 \text{ Golden Globe Wins} \quad \textbf{(11.5)}$$

Let us consider $\hat{p} = 0.5$ as the cutoff value for predicting a movie as a winner of the best picture Academy award. If $\hat{p} \geq 0.5$, then $\ln(\hat{p}/(1-\hat{p})) \geq 0$. Conversely, if $\hat{p} < 0.5$, then $\ln(\hat{p}/(1-\hat{p})) < 0$. Thus, by setting equation (11.5) equal to zero, we can obtain the **decision boundary** for classifying a movie. The decision boundary for a binary classification method separates its predictions of the two different classes. The decision boundary corresponding to a logistic regression classifier is a linear function of the input variables. For these data, the decision boundary of the estimated logistic regression model corresponds to the line:

$$0.458 \text{ Oscar Nominations} + 0.762 \text{ Golden Globe Wins} = 6.44 \quad \textbf{(11.6)}$$

Figure 11.8 illustrates the decision boundary corresponding to equation (11.6). The blue data points correspond to movies that were best picture winners. The orange data points correspond to movies that did not win the best picture. In Figure 11.8, the orange-shaded region to the left of the decision boundary corresponds to the observations for which equation (11.5) results in a negative value of the logged odds (and thus $\hat{p} < 0.5$) so the logistic regression model will classify these observations as not winning (losing). The blue-shaded region to the right of the decision boundary corresponds to the observations for which equation (11.5) results in a positive value of the logged odds (and thus $\hat{p} > 0.5$) so the logistic regression model will classify these observations as winning.

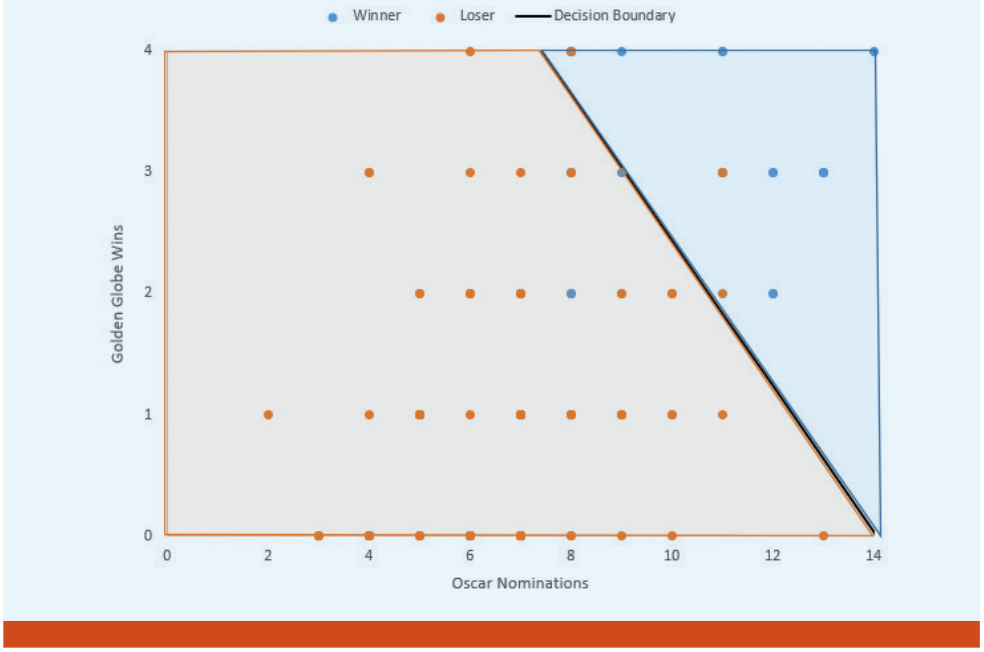
Figure 11.8 visually depicts the logistic regression model's classification performance. A blue data point in the blue-shaded region corresponds to a winning movie which the logistic regression correctly classifies as a winner. A blue data point in the orange-shaded region corresponds to a winning movie which the logistic regression incorrectly classifies as a loser. An orange data point in the orange-shaded region corresponds to a losing movie which the logistic regression correctly classifies as a loser. An orange data point in the blue-shaded region corresponds to a losing movie which the logistic regression incorrectly classifies as a winner.

Table 11.4 Logistic Regression Predictions on Four Validation Set Observations in Oscars Example

Total Number of Oscar Nominations	Predicted Probability of Winning	Predicted Class	Actual Class
14	0.89	Winner	Winner
11	0.58	Winner	Loser
10	0.44	Loser	Loser
6	0.07	Loser	Winner

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.8 Decision Boundary for Logistic Regression Model in Oscars Example



Notes + Comments

As with multiple linear regression, strong collinearity between the independent variables $x_1, x_2, \ldots, x_q$ in a logistic regression model can distort the estimation of the coefficients $b_1, b_2, \ldots, b_q$ in equation (11.1). If we are constructing a logistic regression model to explain and quantify a relationship between the set of independent variables and the log odds of an event occurring, then it is recommended to avoid models that include independent variables that are highly correlated. However, if the purpose of a logistic regression model is to classify observations, multicollinearity does not affect predictive capability so correlated independent variables are not a concern and the model should be evaluated based on its classification performance on validation and test sets.

11.4 *k*-Nearest Neighbors Classification

For a classification task, the ***k*-nearest neighbor (*k*-NN)** method bases its prediction of an observation's categorical outcome variable on the k most similar observations from the training set, where similarity is measured with respect to the set of input variables (features). In the k-NN method, an observation is classified as Class 1 if the proportion of Class 1 observations in the k-nearest neighbors in the training set is greater than or equal to a specified cutoff value (typically 0.5).

Unlike logistic regression, which uses a training set to generalize relationships in the data via the logistic equation and then applies this parametric model to estimate the class probabilities of observations in a validation set, a nearest-neighbor classifier performs no calculations on the training set prior to being supplied an observation to predict. Instead, k-NN is called a **lazy learner** because it directly uses the entire training set by identifying the k most similar training set observations for a specified observation in the validation set and uses these k observations to make its classification prediction.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

We discuss Euclidean distance and other distance measures in more detail in Chapter 6.

There are two primary algorithm choices to specify for the k-NN method: how to measure the distance between observations and how many neighbors (k) to consider in the calculation of the predicted outcome value. Data mining software commonly employs Euclidean distance in the k-NN method to measure the similarity between observations, but other distances measures can be used. The value of k can plausibly range from 1 to n, the number of observations in the training set. If $k = 1$, then the classification of a new observation is set to be equal to the class of the single most similar observation from the training set. At the other extreme, if $k = n$, then the new observation's class is assigned to the most common class in the training set. Smaller values of k are more susceptible to noise in the training set, while larger values of k may fail to capture the relationship between the features and classes of the target variable. Values of k from 1 to $\sqrt{n}$ are typically considered. The best value of k can be determined by building models for a range of k values and then selecting the value of $k^\star$ that results in the best classification performance with respect to the measure of choice (e.g., AUC) on the validation set. Note that the use of the validation set to identify $k^\star$ in this manner implies that the method should be applied to a test set with this value of $k^\star$ to accurately estimate the classification performance on future data.

To illustrate, suppose that a training set consists of the 10 observations listed in Table 11.5. For this example, we will refer to an observation with Loan Default $= 1$ as a Class 1 observation and an observation with Loan Default $= 0$ as a Class 0 observation. Consider a validation set observation with Average Balance $= 900$, Age $= 28$, and Loan Default $= 0$. The k-NN method seeks to classify this observation based on the k most similar training set observations with respect to the values of Average Balance and Age.

In Chapter 2, we discuss z-scores.

Before computing the similarity between a validation set observation and the observations in the training set, it is common practice to standardize the values of all variables. By replacing the original values of each variable with the corresponding z-score, we avoid the computation of Euclidean distance being disproportionately affected by the scale of the variables. The average value of the Average Balance variable in the training set is 1,285 and the standard deviation is 2,029. The average and standard deviation of the Age variable are 38.2 and 10.2, respectively. By calculating the z-score, Observation 1's standardized value of Average Balance is $(49 - 1{,}285)/2{,}029 = -0.609$ and its standardized value of Age is $(38 - 38.2)/10.2 = -0.02$. The validation set observation with Average Balance $= 900$ and Age $= 28$ is also standardized with respect to the training set, resulting in a standardized Average Balance value of $(900 - 1{,}285)/2{,}209 = -0.190$ and a standardized Age value of $(28 - 38.2)/10.2 = -1.003$.

Table 11.5 Training Set for k-NN Classifier

Observation	Average Balance	Age	Average Balance (standardized)	Age (standardized)	Loan Default
1	49	38	−0.609	−0.020	1
2	671	26	−0.303	−1.199	1
3	772	47	−0.253	0.865	1
4	136	48	−0.566	0.963	1
5	123	40	−0.573	0.177	1
6	36	29	−0.616	−0.904	0
7	192	31	−0.539	−0.708	0
8	6,574	35	2.606	−0.315	0
9	2,200	58	0.451	1.946	0
10	2,100	30	0.401	−0.806	0
$\bar{x}$:	1,285	38.2	0	0	
s:	2,029	10.2	1	1	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.9 Classifying a Validation Set Observation with *k*-NN

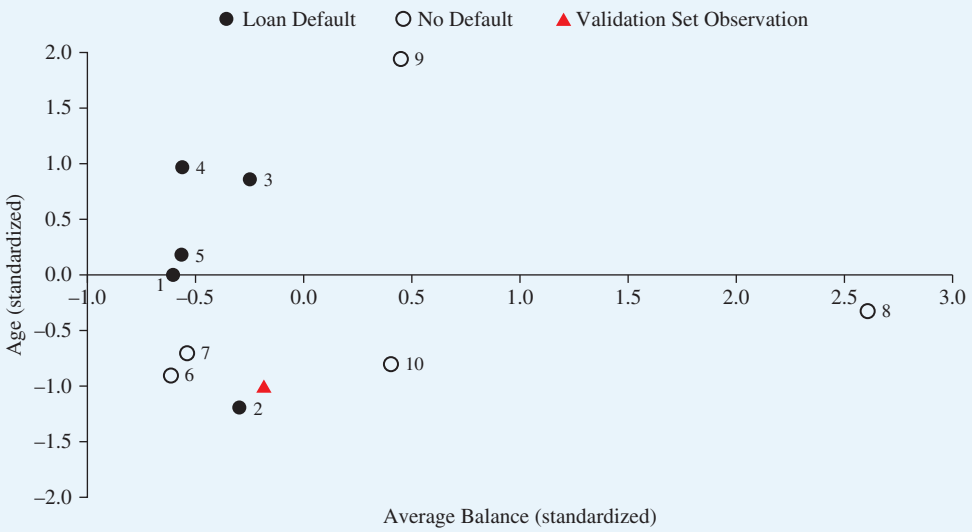


Figure 11.9 shows a plot of the 10 training-set observations and the validation set observation to be classified according to their standardized variable values. To classify the new observation, we will use a cutoff value of 0.5. Consider classifying the validation set observation using its $k = 3$ nearest neighbors in the training set. From Figure 11.9, we see that Observations 2, 6, and 7 are three most similar observations to the validation set observation. Because 1/3 (< 0.5) of these observations are Loan Default $= 1$, the *k*-NN model with $k = 3$ classifies this validation set observation as Loan Default $= 0$. Table 11.6 summarizes the classification of the validation set observation for values of *k* ranging from 1 to 10.

By classifying an entire validation set of observations with various values of *k*, the best value of *k* with respect to the performance measure of choice can be identified. Suppose that after a cross-validation experiment, the value $k = 5$ was identified as resulting in good predictive performance. Based on the 10 training set observations in Table 11.5,

Table 11.6 Classification of Observation with Average Balance = 900 and Age = 28 for Different Values of *k*

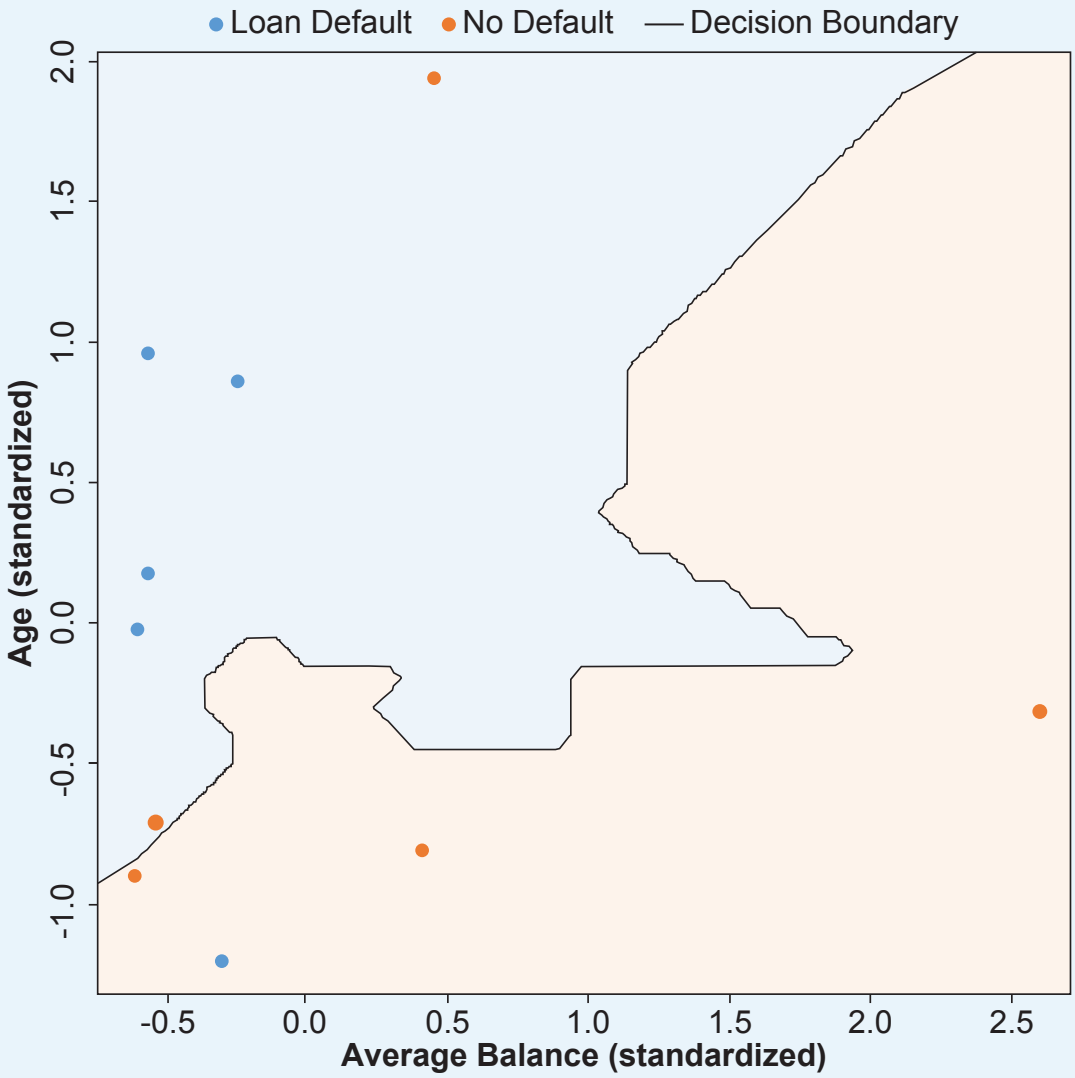
k	Set of Neighbors	% of Class 1 Neighbors	Classification
1	2	1.00	1
2	2,6	0.50	1
3	2,6,7	0.33	0
4	2,6,7,10	0.25	0
5	2,6,7,10,1	0.40	0
6	2,6,7,10,1,5	0.50	1
7	2,6,7,10,1,5,3	0.57	1
8	2,6,7,10,1,5,3,4	0.63	1
9	2,6,7,10,1,5,3,4,8	0.56	1
10	2,6,7,10,1,5,3,4,8,9	0.50	1

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.10 displays the decision boundary for the k-NN model based on $k = 5$. The blue data points correspond to individuals who defaulted on a loan. The orange data points correspond to individuals who did not default on a loan. In Figure 11.10, the orange-shaded region to the right of the decision boundary corresponds to observations for which less than half of the $k = 5$ nearest training set observations correspond to Loan Default = 1, so the k-NN model will classify these observations as not defaulting. The blue-shaded region to the left of the decision boundary corresponds to observations for which more than half of the $k = 5$ nearest training set observations correspond to Loan Default = 1, so the k-NN model will classify these observations as defaulting.

From Figure 11.10, we observe that the shape of the decision boundary corresponding to a k-NN model can be quite irregular. Unlike logistic regression, which creates a linear decision boundary with respect to the input features (see Figure 11.8), the shape of the decision boundary for k-NN is generally not linear and is directly shaped by the observations in the test set and the specified value of k.

Figure 11.10 k-NN Decision Boundary for $k = 5$



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

11.5 Classification Trees

Because there are several algorithm variants for constructing a classification tree, various data mining software packages may obtain different trees. In this section, we present a general description of tree-based logic which may slightly differ from the implementation of classification trees in a specific data mining software.

A **classification tree** is a predictive model consisting of a series of rules on the input features so that each successive rule splits the training set of observations into smaller subsets that are more homogeneous with respect to the binary outcome variable. In this section, we begin by describing how a classification tree is constructed based on a training set. Then we discuss using the classification tree to make predictions and pruning a classification tree to prevent overfitting.

Constructing a Classification Tree

At each iteration in the construction of a classification tree, a subset of observations is further divided into two new subsets using a threshold rule based on the observations' values of a single input variable. The input variable selected for the rule and the threshold value of this input variable upon which to divide the observations is determined to decrease the aggregate **impurity** of the binary outcome variable in the two resulting subsets as much as possible. Impurity is a measure of the heterogeneity in a group of observations' outcome classes. Thus, the sequence of rules in a classification tree can be thought of a series of questions that successively split observations into smaller groups, so that the resulting groups contain an ever-increasing majority of one class.

DATA*file*
hhi

To demonstrate the classification tree method, we consider an example involving Hawaiian Ham Inc. (HHI), a company that specializes in the development of software that filters out unwanted e-mail messages, often referred to as "spam." The file *hhi* contains a sample of data that HHI has collected. For 4,601 e-mail messages, HHI has collected whether the message was "spam" (Class 1) or "not spam" (Class 0), as well as the frequency of the "!" character and the "$" character, expressed as a percentage (between 0 and 100) of characters in the message. In the file *hhi*, the variable EmailID is an email identifier, the variable Exclamation corresponds to the percentage of the "!" character in an email, the variable Dollar corresponds to the percentage of the "$" character in an email, and the variable Spam is "Yes" if the email was spam and "No" if the variable was not spam.

To illustrate how a classification tree is constructed, we consider the first 46 observations in the file *hhi* as a training set. In our description of the construction of a classification tree for the HHI example, we will refer to Figures 11.11 and 11.12. Figure 11.11 displays the full classification tree obtained when applying a set of rules that partitions the set of observations to the greatest extent possible. Figure 11.12 contains a scatter chart that provides a geometric illustration of the full classification tree in Figure 11.11. In Figure 11.12, the coordinates of each data marker correspond to the values of Dollar and Exclamation, while the shape and color of the data marker denote whether it corresponds to a spam or nonspam observation.

Let us consider the classification tree in Figure 11.11. The top node in the classification tree of Figure 11.11 is called the **root node**. The root node in Figure 11.11 conveys that in the entire training set there are 46 observations, 26 of which (56.5%) correspond to Spam = No. At each step, the tree logic identifies the split of the variable that results in the least impurity in the two resulting categories.

The first partition is formed by splitting observations into two subsets with respect to the variable Dollar. The left branch corresponds to 28 observations with Dollar ≤ 0.053, and the right branch corresponds to 18 observations with Dollar > 0.053. The split on the variable Dollar at the value 0.053 is selected because it results in the two subsets of the original 46 observations with the least impurity. We see how the impurity has been decreased by noting that we started with a group of 46 observations containing 56.5% Spam = No and split this into a subset of 28 observations with 25/28 = 89.3% Spam = No and a subset of 18 observations with 17/18 = 94.4% Spam = Yes.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.11 Construction Sequence of Branches in a Full Classification Tree

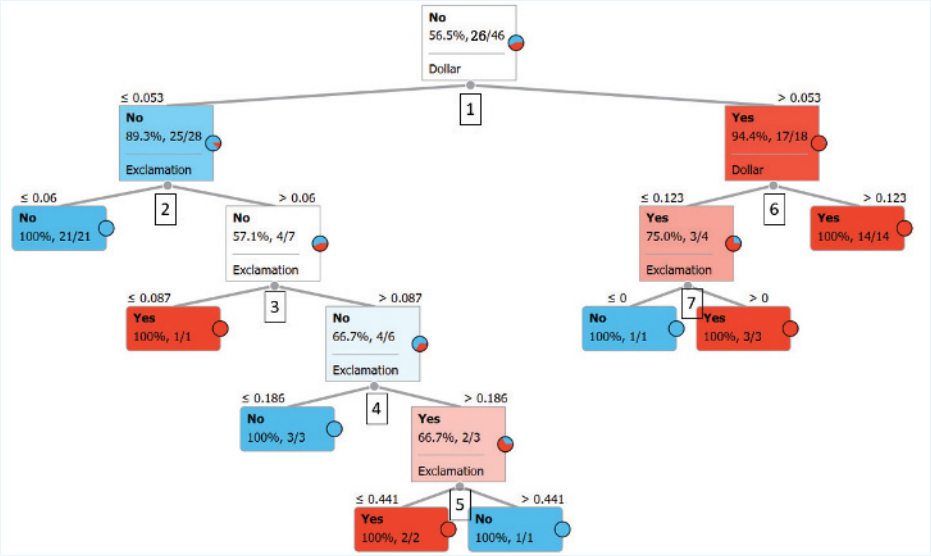
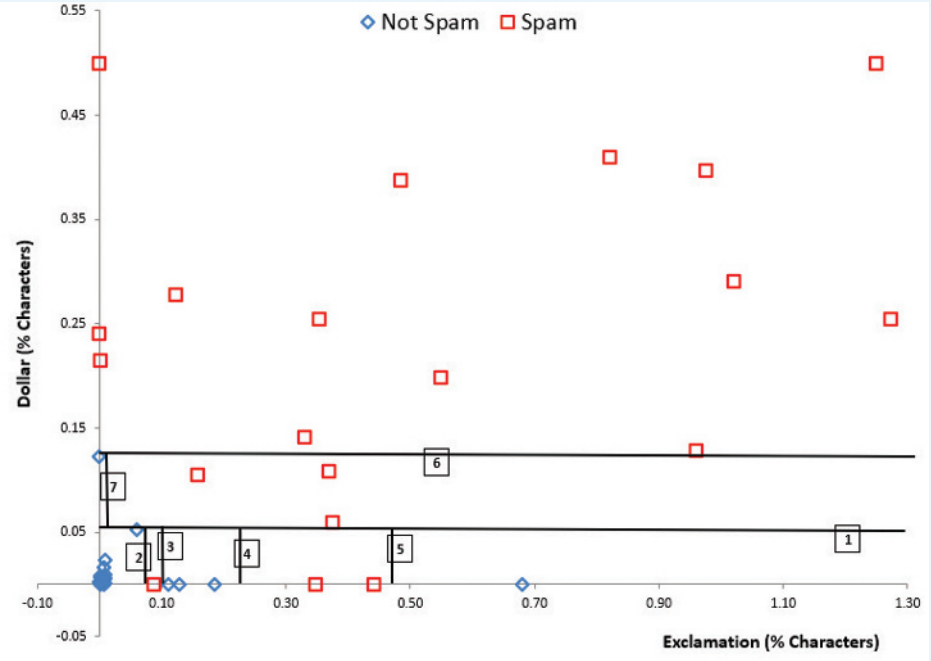


Figure 11.12 is based on all 46 observations, but only 28 of these observations are distinct. Of the 46 observations, 18 of them are not spam and have coordinates (0,0). Another two of observations are spam and have coordinates (0, 0.210).

DATA*file*
trainhhi

Figure 11.12 Geometric Illustration of Full Classification Tree Partitions



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The splitting process is then repeated on these two newly created subsets of observations in a manner that again results in additional subsets with the least impurity. In this tree, the second split is applied to the group of 28 observations with Dollar ≤ 0.053 using the variable Exclamation. The left branch corresponds to 21 observations with Exclamation ≤ 0.06 and the right branch corresponds to 7 observations with Exclamation > 0.06. The 21 observations with Exclamation ≤ 0.06 are 21/21 = 100% Spam = No and the 7 observations with Exclamation > 0.06 are 4/7 = 57.1% Spam = No. No further partitioning of the 21-observation subset with values of Dollar ≤ 0.053 and Exclamation ≤ 0.06 is necessary because there is subset that has zero impurity. A subset of observations that is not split any further is a called a **leaf node**.

At this point in the tree construction process, there are two subsets to further partition. There is a 7-observation subset with values of Dollar ≤ 0.053 and Exclamation > 0.06 on the left side of Figure 11.11 as well as an 18-observation subset with values of Dollar > 0.053 on the right side of Figure 11.11. The partitioning process continues in the order as denoted by the boxed numbers in Figure 11.11. For example, the third partitioning step further splits the 7-observation subset with values of Dollar ≤ 0.053 and Exclamation > 0.06 on the variable Exclamation and the value 0.087. The left branch corresponds to one observation with Exclamation ≤ 0.087 and the right branch corresponds to six observations with Exclamation > 0.087.

The recursive partitioning process for constructing a full classification tree continues until impurity of the subsets of observations with respect to the categorical outcome variable Spam can no longer be reduced by a rule on the value of one of the input variables. In this example, Figure 11.11 shows this occurs when the leaf nodes each contain 100% Spam = Yes or 100% Spam = No. In general, non-zero impurity may occur in a leaf node when multiple observations have the same set of values of the input variables, but have different values of the categorical outcome variable. For this example, that would correspond to a case in which two emails have the same values of Dollar and Exclamation, respectively, but one is spam and the other is not spam.

Figure 11.12 illustrates that when considering only two input variables, a classification tree partitions the observations into rectangular regions. Each variable splitting rule corresponds to a horizontal line (for rules on Dollar) or a vertical line (for rules on Exclamation) and are labeled in the order they are constructed. These rules divide the variable space into eight rectangles, each corresponding to one of the eight leaf nodes in the tree in Figure 11.11.

As Figures 11.11 and 11.12 show, for these data the full classification tree partitions the training set to the extent that each subset contains either all Class 1 observations or all Class 0 observations. In other words, enough decomposition of these data results in a set of leaf nodes with zero impurity, and there are no misclassifications of the training set by this full tree. In general, unless there exist observations that have identical values of all the input variables but different outcome classes, the leaf nodes of the full classification tree will have zero impurity.

Generating Predictions with a Classification Tree

After a final classification tree is constructed from a training set, it can be used to predict observations in a validation set. To classify a validation set observation, the tree's rules are applied to the observation's input features to determine which of the tree's leaf nodes contains the observation. The proportion of Class 1 observations in this leaf node is then compared to the cutoff value to determine the predicted outcome class of the validation set observation. For example, if the cutoff value is 0.5, the majority class of the leaf node containing the validation set observation is the predicted outcome class.

Returning to the HHI example, consider a validation set observation with Exclamation = 0.08, Dollar = 0, and Spam = Yes. To classify this observation with the full classification tree in Figure 11.11, we start at the root node (labeled with a 1) and then travel down the left branch because Dollar = $0 \leq 0.053$. At the next node (labeled with a 2), we travel

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

down the right branch because Exclamation = 0.08 > 0.06. Then, at the node labeled with a 3, we arrive at the leaf node on the left branch because Exclamation = 0.08 ≤ 0.087. Because 1/1 = 100% > 50% of this leaf node's observations are Spam = Yes, we classify (correctly in this case) the validation set observation as Spam = Yes.

To facilitate our explanation of how a classification tree is constructed, we split the 4,601 observations into a 46-observation training set and a validation set with 4,555 observations. In practice, a k-folds cross-validation procedure would be applied using the 4,601 observations.

To assess the predictive performance of the full classification tree, we repeat the classification calculations for an entire validation set. In this case, the 4,555 observations with EmailID = 47 through EmailID = 4,601 in the file *hhi* compose the validation set. Applying the full classification tree to this entire validation set results in a classification error of (990/4555) = 0.217 and an AUC of 0.758.

As we discussed in the previous section, Figures 11.11 and 11.12 show that the full classification tree perfectly characterizes the training set. However, the results from applying the full classification tree to a validation set show that the predictive performance is far from perfect. This is an example of overfitting–the extensive rules of the full classification tree are capturing randomness in the training data rather than generalizable relationships between the input variables and the outcome variable. This suggests that a classification tree with fewer branches may produce more accurate out-of-sample predictions.

The process of obtaining a classification tree with fewer branches than a full classification tree is called **pruning**. There are several ways to prune a full classification tree to obtain a final classification tree with the best possible out-of-sample predictive accuracy. One approach is to sequentially remove branches from the bottom of the tree and assess the out-of-sample predictive performance of the sequence of increasingly smaller classification trees. Starting from the full classification tree in Figure 11.11 with eight leaf nodes, this would result in testing the predictive accuracy of a classification tree with seven leaf nodes, then six leaf nodes, …, and finally only one leaf node (the root node).

Another approach to reducing the size of a classification tree is to specify restrictions on how much a rule can split up a group of observations. One common restriction is to specify that each of the two subgroups resulting from a variable splitting rule must have at least a specified number of observations. Another common restriction is to prohibit the splitting of a group of observations unless it has more than a specified number of observations.

Regardless of the approach used to prune a classification tree, the appropriate size of the tree is determined through assessing its out-of-sample predictive accuracy. For the 46-observation training set in the HHI example, the pruned classification tree in Figure 11.13 achieves the best performance on the validation set. This pruned classification tree achieves a classification error of (974/4555) = 0.208 and an AUC of 0.862 on the 4,555 observations of the validation set compared to the classification error of 0.208 and an AUC of 0.785 for the full classification tree.

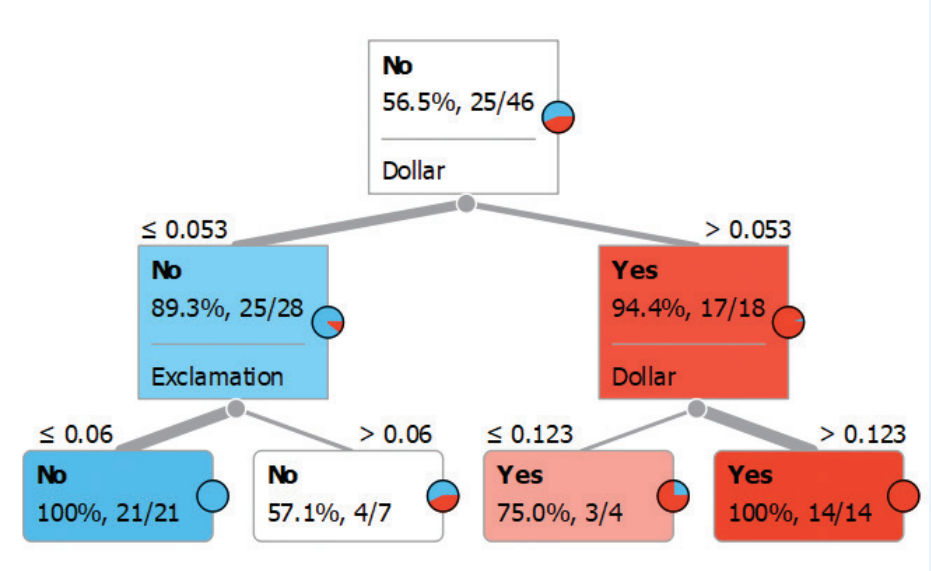
Ensemble Methods

In the previous section, we demonstrated the classification of an observation based on the decision rules of a single classification tree. In this section, we discuss the notion of ensemble methods. In an **ensemble method**, predictions are made based on the combination of a collection of models. For example, instead of basing the classification of a new observation on an individual classification tree, an ensemble method generates a collection of different classification trees and then predicts the class of an observation based on the collective voting of this collection.

To gain an intuitive understanding of why an ensemble of classification models may outperform, on average, any single classification model, consider the task of predicting whether the value of the S&P 500 Index one year from now will be higher or lower than

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.13 Pruned Classification Tree



its current value. Suppose there are 100 financial analysts independently developing their own forecast based on a variety of information. One year from now, there may be one or more analysts who correctly predicted the direction of the S&P Index. However, knowing beforehand which of these analysts would be correct may be difficult to discern. Therefore, instead of trying to pick one of the analysts and depending solely on their forecast, an ensemble approach would combine their forecasts and use this composite prediction to classify whether the S&P Index will go up or down. The two conditions for an ensemble to perform better than a single model are as follows: (1) The individual base models are constructed independently of each other and (2) the individual models perform better than just randomly guessing.

There are two primary steps to an ensemble approach: (1) the construction of a committee of individual base models and (2) the combination of the individual base models' predictions to form a composite prediction. While an ensemble can be composed of any type of individual classification model, the ensemble approach works better with unstable learners. A classification method is an **unstable learner** if relatively small changes in the training set cause its predictions to fluctuate substantially. In this section, we discuss ensemble methods using classification trees, which are known to be unstable. Specifically, we discuss three different ways to construct an ensemble of classification trees: bagging, boosting, and random forests.

In the **bagging** approach, the committee of individual base models is generated by first constructing multiple training sets by repeated random sampling of the n observations in the original data *with replacement*. Because the sampling is done with replacement, this process may result in training sets which contain multiple copies of some observations and no copies of other observations. The bagging approach then trains a classification model on each of the m training sets and generates the ensemble prediction based on the majority of the m individual predictions. A basic premise of bagging is that training

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 11.7 Original Training Set of 10 Observations

Age	29	31	35	38	47	48	53	54	58	70
Loan default	0	0	0	1	1	1	1	0	0	0

unstable learners on the slightly different *m* training sets will generate diversity in the resulting predictions.

To demonstrate bagging, we consider the task of classifying customers as defaulting or not defaulting on their loan, using only their age. Table 11.7 contains the 10 observations in the original training data. Table 11.8 shows the results of generating 10 new training sets by randomly sampling from the original data with replacement. For each of these training sets, we construct a one-rule classification tree (called a **stump**) that minimizes the impurity of the resulting partition. The two groups of each training sample are illustrated with a vertical red line and accompanying decision rule.

Table 11.9 shows the results of applying this ensemble of 10 classification trees to a validation set consisting of 10 observations. The ensemble method bases its classification on the aggregate vote of the 10 individual classifications trees; if at least half of the individual trees classify an observation as Class 1, so does the ensemble. Note from Table 11.9 that the 20% classification error rate of the ensemble is lower than any of the individual trees, illustrating the potential advantage of using ensemble methods.

As an alternative to using a separate validation set (as the one in Table 11.9), the predictive performance of the bagging ensemble constructed in Table 11.8 can be assessed by using **out-of-bag estimation**. Out-of-bag estimation leverages the fact that because the training sets for each base model used in the ensemble is trained by sampling the data with replacement, each of these training sets will each be missing an average of about 36.8% of the distinct observations.[1] Each of these missing observations can be used as holdout data to evaluate the subset of models which do not contain the missing observations.

To demonstrate out-of-bag estimation, consider the bagging ensemble in Table 11.8. Because the training sets for each respective tree in Table 11.8 are constructed by sampling the original 10 observations in Table 11.7 with replacement, these training sets do not contain all 10 original observations. For example, the training set for Tree 1 does not contain the (Age, Loan Default) observations of (53, 1); (54, 0); and (70, 0). Therefore, these three observations constitute holdout data for Tree 1. For each tree in the ensemble method, Table 11.10 shows how the tree classifies the observations that were missing from the respective tree's training set. For example, Tree 1's rule is to classify all observations with Age $\leq$ 36.5 as Class 0 and all observations with Age $>$ 36.5 as Class 1. Therefore, it classifies the three observations (Age = 53, Age = 54, and Age = 70) as Class 1.

After each tree predicts the observations missing from its respective training set, the classification of each of the original 10 observations is executed by aggregating the votes of the trees participating on each respective observation and classifying the observation as the majority class. For example, observation (31, 0) is classified as Class 1 by Tree 7, as Class 1 by Tree 9, and as Class 0 by Tree 10; therefore, the out-of-bag classification for this observation is Class 1 by majority vote. Based on this out-of-bag classification, an out-of-bag classification error can be computed. From Table 11.10, we see that the

[1]For a data set with *n* observations, the probability of an observation not being selected when randomly sampling the data set with replacement is $\left(\frac{n-1}{n}\right)^n$. For large *n*, $\left(\frac{n-1}{n}\right)^n$ is approximately 0.368.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 11.8 Bagging: Generation of 10 New Training Sets and Corresponding Classification Trees

Sample 1/Tree 1					Age ≤ 36.5					
Age	29	31	31	35	38	38	47	48	58	58
Loan default	0	0	0	0	1	1	1	1	0	0
Prediction	0	0	0	0	1	1	1	1	1	1
Sample 2/Tree 2					Age ≤ 50.5					
Age	29	31	35	38	47	54	58	70	70	70
Loan default	0	0	0	1	1	0	0	0	0	0
Prediction	0	0	0	0	0	0	0	0	0	0
Sample 3/Tree 3					Age ≤ 36.5					
Age	29	31	35	38	38	47	53	53	54	58
Loan default	0	0	0	1	1	1	1	1	0	0
Prediction	0	0	0	1	1	1	1	1	1	1
Sample 4/Tree 4					Age ≤ 34.5					
Age	29	29	31	38	38	47	47	53	54	58
Loan default	0	0	0	1	1	1	1	1	0	0
Prediction	0	0	0	1	1	1	1	1	1	1
Sample 5/Tree 5					Age ≤ 39					
Age	29	29	31	47	48	48	48	70	70	70
Loan default	0	0	0	1	1	1	1	0	0	0
Prediction	0	0	0	1	1	1	1	1	1	1
Sample 6/Tree 6					Age ≤ 53.5					
Age	31	38	47	48	53	53	53	54	58	70
Loan default	0	1	1	1	1	1	1	0	0	0
Prediction	1	1	1	1	1	1	1	0	0	0
Sample 7/Tree 7					Age ≤ 53.5					
Age	29	38	38	48	53	54	58	58	58	70
Loan default	0	1	1	1	1	0	0	0	0	0
Prediction	1	1	1	1	1	0	0	0	0	0
Sample 8/Tree 8					Age ≤ 53.5					
Age	29	31	47	47	47	53	53	54	58	70
Loan default	0	0	1	1	1	1	1	0	0	0
Prediction	1	1	1	1	1	1	1	0	0	0
Sample 9/Tree 9					Age ≤ 53.5					
Age	29	35	38	38	48	53	53	54	70	70
Loan default	0	0	1	1	1	1	1	0	0	0
Prediction	1	1	1	1	1	1	1	0	0	0
Sample 10/Tree 10					Age ≤ 14.5					
Age	29	29	29	29	35	35	54	54	58	58
Loan default	0	0	0	0	0	0	0	0	0	0
Prediction	0	0	0	0	0	0	0	0	0	0

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 11.9 Classification of 10 Observations from Validation Set with Bagging Ensemble

Age	26	29	30	32	34	37	42	47	48	54	Overall Error Rate (%)
Loan default	1	0	0	0	0	1	0	1	1	0	
Tree 1	0	0	0	0	0	1	1	1	1	1	30
Tree 2	0	0	0	0	0	0	0	0	0	0	40
Tree 3	0	0	0	0	0	1	1	1	1	1	30
Tree 4	0	0	0	0	0	1	1	1	1	1	30
Tree 5	0	0	0	0	0	0	1	1	1	1	40
Tree 6	1	1	1	1	1	1	1	1	1	0	50
Tree 7	1	1	1	1	1	1	1	1	1	0	50
Tree 8	1	1	1	1	1	1	1	1	1	0	50
Tree 9	1	1	1	1	1	1	1	1	1	0	50
Tree 10	0	0	0	0	0	0	0	0	0	0	40
Average Vote	0.4	0.4	0.4	0.4	0.4	0.7	0.8	0.8	0.8	0.4	
Bagging Ensemble	0	0	0	0	0	1	1	1	1	0	20

out-of-bag classification is incorrect on 7 of the 10 observations resulting in a 70% overall error rate. While the out-of-bag estimate of the overall error rate is much higher than the 20% overall error rate observed on the validation set in Table 11.9, these values are typically much closer for larger data sets. However, it is important to note that out-of-bag estimation does not assess performance as thoroughly as a separate validation set or a cross-validation procedure. In general, out-of-bag estimates will be more conservative

Table 11.10 Out-of-Bag Classifications of Observations Missing from Each Tree's Training Set

Age	29	31	35	38	47	48	53	54	58	70
Loan default	0	0	0	1	1	1	1	0	0	0
Tree 1							1	1		1
Tree 2						0	0			
Tree 3						1				1
Tree 4			1			1				1
Tree 5			0	0			1	1	1	
Tree 6	1		1							
Tree 7		1	1		1					
Tree 8			1	1		1				
Tree 9		1			1				1	
Tree 10		0		0	0	0	0			0
Aggregate Vote	1/1	2/3	4/5	1/3	2/3	3/5	2/4	2/2	2/2	3/4
Out-of-Bag Classification	1	1	1	0	1	1	1	1	1	1

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

than the analogous performance measures based on a validation set. That is, an out-of-bag error estimate will be larger than a validation-based error estimate, and an out-of-bag AUC estimate will be smaller than a validation-based AUC estimate. Looking down the columns of Table 11.10, we observe that the aggregate vote of each observation is based on a different subset of the trees.[2] Looking across the rows of Table 11.10, we observe that the trees participate in varying numbers of aggregate votes.[3] In a true validation set, each observation would be predicted by each tree in the ensemble. As illustrated in Table 11.10, between two to five trees participate in the vote of the 10 original observations. However, in Table 11.9, each of the 10 observations in a validation set is predicted by all 10 trees from the ensemble.

Similar to bagging, the **boosting** method generates its committee of individual classifiers by sampling multiple training sets. However, boosting differs from bagging in how it samples the multiple training sets and how it weights the resulting classification models to compute the ensemble's prediction. Unlike bagging, in which the generation of the training sets (and corresponding classifiers) are independent of each other, in boosting the construction of the jth training set (and the corresponding classifier) is dependent on the previous $j-1$ training sets and classifiers. Boosting iteratively adapts the construction of the jth training set based on the prediction error of the classifiers constructed on the previous $j-1$ training sets. There are several different implementations of boosting, but we describe one common one here. In this implementation, to generate the first training set, each of the n observations in the original data is initially given equal weight of being selected. That is, each observation i has weight $w_i = 1/n$. A classification model is then trained on this training set and is used to predict the outcome of the n observations in the original data. The weight of each observation i is then adjusted based on the degree of its prediction error. For example, in a classification problem, if an observation i is misclassified by a classifier, then its weight w_i is increased, but if it is correctly classified, then its weight w_i is decreased. The next training set is then generated by sampling the observations according to the updated weights. In this manner, the next training set is more likely to contain observations that have been mispredicted in early iterations. However, this dependence also means that it is possible for a boosting ensemble to become overfit if too many individual classifiers are constructed.

To combine the predictions of the m individual models from the m training sets, boosting weights the vote of each individual model based on its overall prediction error. For example, suppose that the classifier associated with the jth training set has a large prediction error and the classifier associated with the kth training set has a small prediction error. Then the classification votes of the jth classifier will be weighted less than the classification votes of the kth classifier when they are combined. Note that this method differs from bagging, in which each of the individual classifiers has an equally weighted vote.

A **random forest** is an extension of bagging specifically tailored for use with classification trees. As in bagging, the random forests approach generates multiple training sets by randomly sampling with replacement the n observations in the original data. However, when constructing a tree model for each separate training set, each tree is restricted to using only a specified number of randomly selected input variables. For example, suppose we are attempting to classify a tax return as fraudulent or not fraudulent and there are q input variables. For each of the m generated training sets, an individual classification tree is constructed based on splitting rules based on f randomly selected input variables, where f is smaller than q. The individual classification trees are referred to as weak learners because

[2]It is possible that an observation is included in every base model's training set. In this case, the observation would not contribute to the out-of-bag estimation process.

[3]It is possible, although unlikely, that a tree's training set contains every distinct observation and therefore has no original observations missing from its training set. In this case, the tree would not contribute to the out-of-bag estimation process.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

they are only allowed to consider a small subset of input variables. A **weak learner** is a predictive model whose individual predictions are only slightly better than random guessing. Ensembles commonly consist of unstable learners that are intentionally weakened in an attempt to improve the independence of their predictions. Pruning the individual trees in a random forest is not typically necessary due to their restricted choice of input variables and their incorporation into an ensemble reduces the likelihood of overfitting. While the best number of individual trees in the random forest depends on the data, it is not unusual for a random forest to consist of hundreds and even thousands of individual trees.

For most data, the predictive performance of boosting ensembles exceeds the predictive performance of bagging ensembles. Boosting achieves its performance advantage because (1) it evolves its committee of models by focusing on observations that are mispredicted, and (2) the member models' votes are weighted by their accuracy. However, boosting is more computationally expensive than bagging. Because there is no adaptive feedback in a bagging approach, all m training sets and corresponding models can be implemented simultaneously. However, in boosting, the first training set and predictive model guide the construction of the second training set and predictive model, and so on. The random forests approach has performance similar to boosting, but maintains the computational simplicity of bagging.

Notes + Comments

1. The Gini index and entropy are two common ways to measure the impurity of a node in a classification tree. For a node of n observations, k of which are Class 1 and $n - k$ of which are Class 0, these two measures are computed as:

$$\text{Gini index} = 1 - \left(\frac{k}{n}\right)^2 - \left(\frac{n-k}{n}\right)^2$$

$$\text{entropy} = -\left(\frac{k}{n}\right)\log_2\left(\frac{k}{n}\right) - \left(\frac{n-k}{n}\right)\log_2\left(\frac{n-k}{n}\right)$$

The shapes of these two functions are very similar, with both measures reflecting a maximum level of impurity when 50% of a node's observations are Class 1, and a minimum level of impurity when 0% (or 100%) of a node's observations are Class 1.

For the root node in Figure 11.11,

$$\text{root node entropy} = -\left(\frac{20}{46}\right)\log_2\left(\frac{20}{46}\right) - \left(\frac{26}{46}\right)\log_2\left(\frac{26}{46}\right) = 0.988$$

Splitting the root node into 28 observations with Dollar ≤ 0.053 and 18 observations with Dollar > 0.053 results in

$$\text{left branch entropy} = -\left(\frac{3}{28}\right)\log_2\left(\frac{3}{28}\right) - \left(\frac{25}{28}\right)\log_2\left(\frac{25}{28}\right) = 0.491$$

$$\text{right branch entropy} = -\left(\frac{17}{18}\right)\log_2\left(\frac{17}{18}\right) - \left(\frac{1}{18}\right)\log_2\left(\frac{1}{18}\right) = 0.310$$

The aggregate entropy of these two subsets is (28/46)(0.491) + (18/46)(0.310) = 0.420. Thus, partitioning the root node of 46 observations into these two subsets decreased node impurity as measured by entropy by 0.988 − 0.420 = 0.568.

2. In this section's examples of classification trees, the input variables are quantitative. When considering a categorical input variable, rather than considering possible threshold values, the splitting procedure will consider the possible ways to split the categories of the input variable into two groups.

11.6 Neural Network Classification

A **neural network** is a predictive data mining method whose structure is motivated by the biological functioning of the brain. The relatively simple neural networks we introduce are the foundation of the more sophisticated **deep learning** methods used for applications, such as image recognition and natural language processing. In this section, we describe how a neural network makes predictions about a binary outcome variable, y, by processing values of a set of input variables, $x_1, x_2, \ldots, x_q$, through a series of calculations involving weights and the application of functions. We begin by describing the structure of a neural network and how this structure is used to generate a prediction. Then we describe how a neural network estimates its parameter values from data to model complex relationships between the input features and the outcome variable.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Structure of a Neural Network

A neural network consists of a layered system of nodes called **neurons**. Figure 11.14 illustrates a neural network with four layers: an input layer, two hidden layers, and an output layer. The **input layer** is a set of neurons in which values of the input features enter the neural network. In Figure 11.14, the input layer consists of three neurons, corresponding to the input features x_1, x_2, and x_3. Typically, the number of neurons in the input layer corresponds to the number of input features.

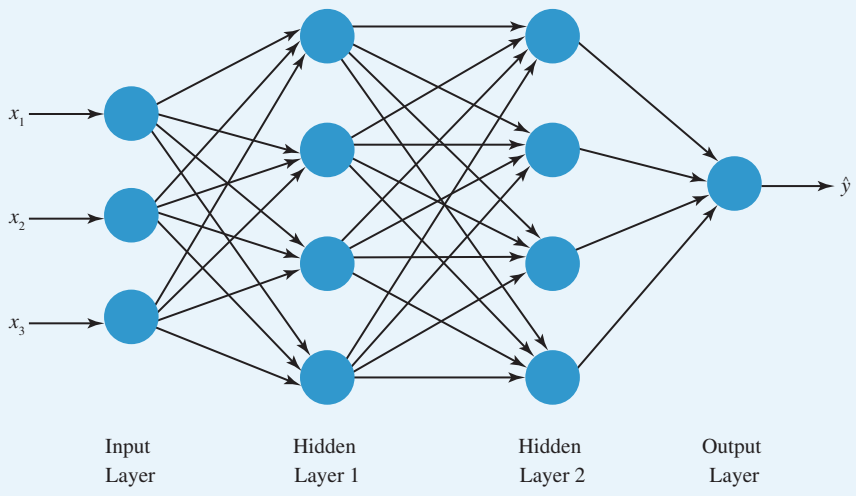
Each layer between the first layer (the input layer) and the last layer (the output layer) is called a **hidden layer**. In the construction of a neural network, the number of hidden layers and the number of neurons in each hidden layer is an important design decision. As the number of hidden layers and number of neurons in each hidden layer increases, the neural network gains the capability of capturing more complex relationships between the input features and the outcome variable. However, the amount of data required to estimate the parameters reflecting these relationships also increases.

The last layer in a neural network is called the **output layer**. The number of neurons in the output layer corresponds to the number of different outcome variables being predicted. For classification tasks we consider in this chapter, the output layer will consist of a single neuron whose output will correspond to the probability an observation's binary categorical outcome is Class 1.

The neural network in Figure 11.14 is called a **feed forward neural network** because the prediction calculations are processed in only one direction—from the input layer in which the values of the input variables enter, through the hidden layer(s) of the network, until the predicted probability of the outcome variable being Class 1 is generated at the output layer. The arrows entering and exiting each neuron in the neural network are known as arcs. We will exclusively consider feed forward neural networks.

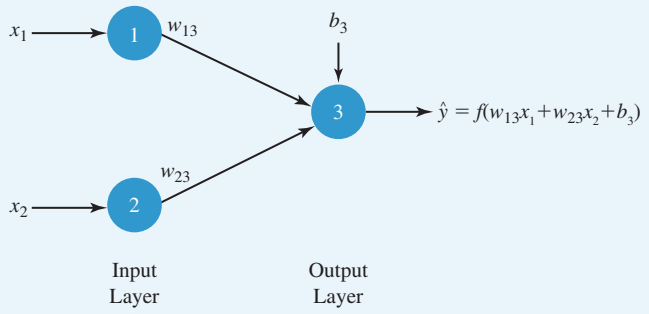
We begin our discussion of how a neural network generates its predictions by considering the simplest form of a neural network. A **perceptron** is a neural network with no hidden layers (only an input layer and output layer). Figure 11.15 contains a perceptron with two input neurons connected to a single output neuron. Figure 11.15 displays the basic components of a neural network: neuron weights (w_{13}, w_{23}), a bias term (b_3), and an activation function (f). In Figure 11.15, w_{13} is the weight applied to the neuron corresponding to feature x_1 and w_{23} is the weight applied to the neuron corresponding to feature x_2.

Figure 11.14 A Feed Forward Neural Network



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.15 Perceptron Model

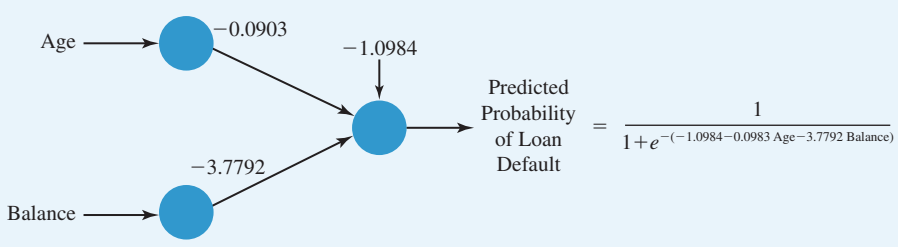


A **neuron weight** represents the strength of a neuron's output signal to a neuron in the next layer. In Figure 11.15, a bias term b_3 is associated with the output neuron. The **bias term** amplifies (if $b_3 > 0$) or dampens (if $b_3 < 0$) the input signal into a neuron. At the output neuron, the predicted value of the outcome variable is obtained by computing $f(w_{13}x_1 + w_{23}x_2 + b_3)$ for a specified activation function f. An **activation function** is a function that produces an output from a neuron based on inputs into a neuron.

Let us consider the application of the perceptron to the Optiva example. Analysts at Optiva Credit Union are interested in predicting whether a customer will default on their loan using information on the customer's age (Age) and average checking account balance (Balance). To compute a prediction using the perceptron in Figure 11.15, we need to specify values of the neuron weights (w_{13}, w_{23}), the bias term (b_3), and the form of the activation function (f). In the next subsection, we describe how a neural network learns values of the neuron weights and bias terms to improve the accuracy of its predictions, but for now, suppose we know these values. Figure 11.16 displays a version of the perception in which $w_{13} = -0.0903$, $w_{23} = -3.7792$, $b_3 = -1.0984$, and the activation function is $f(x) = \frac{1}{1 - e^{-x}}$, the logistic function. As Figure 11.16 shows, implementing a neural network classifier in its simplest architecture with the logistic function as the activation function results in an equation for the predicted probability of a loan default that has the same structural form as the logistic regression model.

The power of a neural network classifier comes from constructing an architecture of one or more hidden layers to learn complex (nonlinear) relationships between the input features and the outcome variable. To demonstrate how a more complicated neural network generates

Figure 11.16 Perceptron Model for Optiva Using Logistic Activation Function



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a prediction, Figure 11.17 displays a neural network with two input variables, one hidden layer of three neurons, and one output variable. Figure 11.17 displays the system of neuron weights, bias terms, and activation functions that the neural network applies in a series of calculations. For every neuron in a layer, there is a weight associated with its connection to each neuron in the next layer. In Figure 11.17, the connection between neuron 1 and neuron 3 has weight w_{13}. In addition, there is a bias term associated with each neuron in the hidden layer and the output layer. In Figure 11.17, the bias term b_4 is associated with neuron 4. At each neuron in the hidden layer and output layer, an activation function is applied to the weighted sum of the inputs and corresponding bias term to generate an output value for the neuron. In Figure 11.17, the output of neuron 5 is produced as $\hat{x}_5 = f(w_{15}x_1 + w_{25}x_2 + b_5)$. The output of the neuron in the output layer corresponds to the predicted probability of the outcome variable being Class 1, $\hat{y} = f(w_{36}\hat{x}_3 + w_{46}\hat{x}_4 + w_{56}\hat{x}_5 + b_6)$.

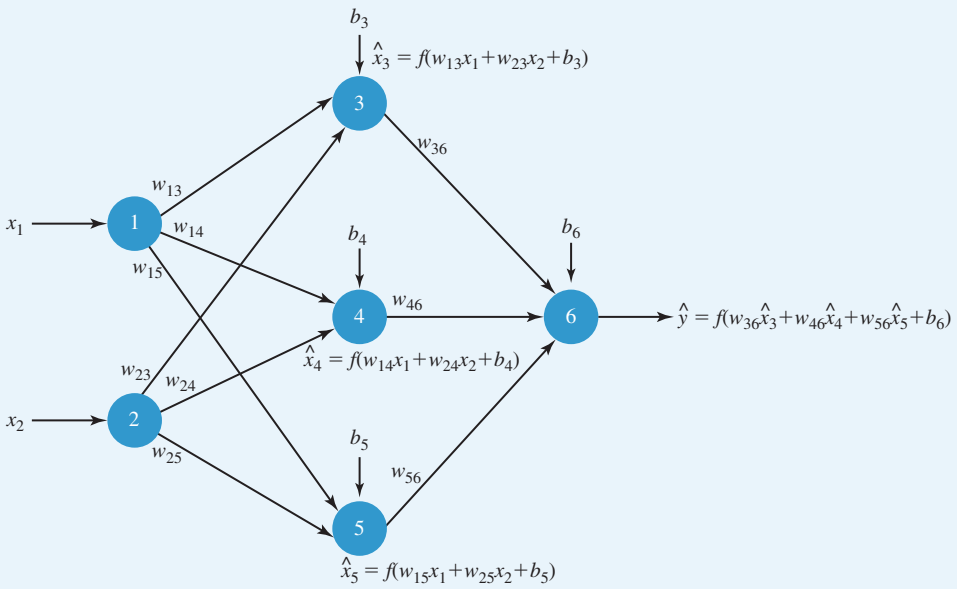
Now let us consider applying the neural network with the structure in Figure 11.17 to predict the probabilities of customers defaulting on their loans in the Optiva example using information on the age of the customer (Age) and their average checking account balance (Balance). To compute a prediction using the neural network, we need values of the neuron weights (w_{13}, w_{14}, w_{15}, w_{23}, …), the bias terms (b_3, b_4, b_5, b_6), and the form of the activation functions (f) at the neurons in the hidden layer and output layer. In the next subsection, we describe how a neural network iteratively learns the good values of the neuron weights and bias terms, but for now we will use the initial values shown in Figure 11.18.

Consider a neural network to predict whether a 38-year old customer with an average checking account balance of $49 will default on their loan. For a neural network, it is common to scale the value of each input feature x to be between 0 and 1. **Min–max scaling** of a variable x is achieved with equation (11.7).

$$x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{11.7}$$

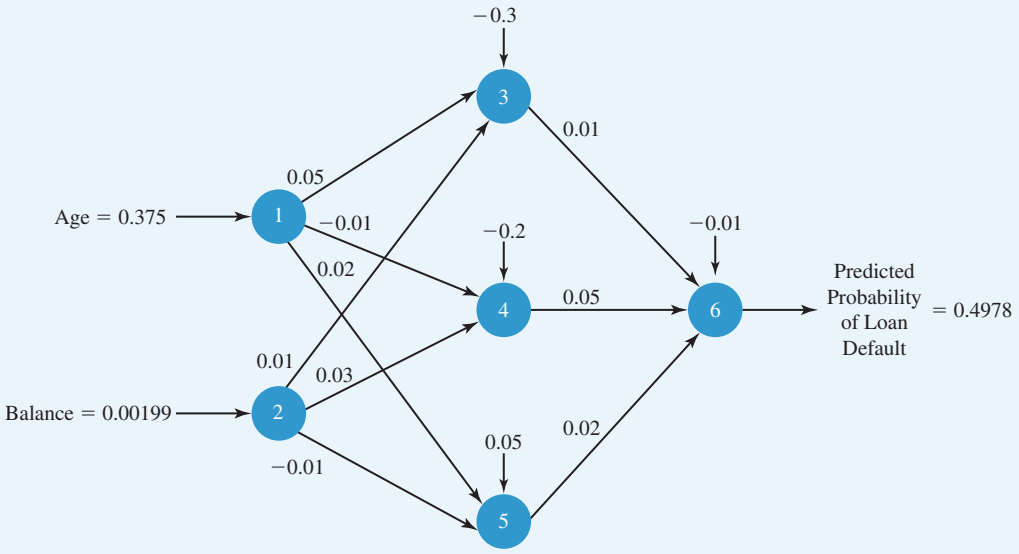
where x_{min} and x_{max} refer to the minimum and maximum observations, respectively, for the variable x.

Figure 11.17 Neuron Weights, Bias Terms, and Activation Functions of a Neural Network



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 11.18 Initial Prediction of Probability of Loan Default from Neural Network



For example, suppose that the minimum and maximum values of Age in the training set used to estimate the parameters of the neural network are 26 and 58 years. Then the scaled value of Age for a 38-year old customer is:

$$\text{Age}_{\text{scaled}} = \frac{\text{Age} - \text{Age}_{\text{min}}}{\text{Age}_{\text{max}} - \text{Age}_{\text{min}}} = \frac{38 - 26}{58 - 26} = 0.375$$

Similarly, suppose that the minimum and maximum values of Balance in the training set used to estimate the parameters of the neural network are \$6,574 and \$36. Then the scaled value of Balance = \$49 is:

$$\text{Balance}_{\text{scaled}} = \frac{\text{Balance} - \text{Balance}_{\text{min}}}{\text{Balance}_{\text{max}} - \text{Balance}_{\text{min}}} = \frac{49 - 36}{6{,}574 - 36} = 0.00199$$

As shown in Figure 11.18, the values of $\text{Age}_{\text{scaled}} = 0.375$ and $\text{Balance}_{\text{scaled}} = 0.00199$ are the values for the input layer.

Given the values of the input features, the output values of the neurons in the next layer can be computed for a specified activation function. Activation functions for neural networks are typically nonlinear functions that allow the neural network to capture complex relationships between the variables. A common activation function in the hidden layers of a neural network is called the **rectified linear unit (ReLu) function**. The ReLu function is

$$f^{\text{ReLu}}(x) = \max\{0, x\}$$

Using the ReLu function for the hidden layer, we obtain the following output values for the neurons in the hidden layer of Figure 11.18:

$$\hat{x}_3 = \max\{0, w_{13}x_1 + w_{23}x_2 + b_3\} = \max\{0, (0.05)(0.375) + (0.01)(0.00199) - 0.3\}$$
$$= \max\{0, -0.281\} = 0$$

$$\hat{x}_4 = \max\{0, w_{14}x_1 + w_{24}x_2 + b_4\} = \max\{0, (-0.01)(0.375) + (0.03)(0.00199) - 0.2\}$$
$$= \max\{0, -0.204\} = 0$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$$\hat{x}_5 = \max\{0, w_{15}x_1 + w_{25}x_2 + b_5\} = \max\{0,(0.02)(0.375) - (0.01)(0.00199) + 0.5\}$$
$$= \max\{0, 0.057\} = 0.057$$

The output values of the three neurons in the output layer then become inputs to the computation of the neuron in the output layer. For a binary classification task, the **logistic function** is a common activation function for the output layer of a neural network. The logistic function is

$$f^{\text{logistic}}(x) = \frac{1}{1+e^{-x}}$$

For the neural network in Figure 11.18, the calculation of the output layer using the logistic activation function is

$$\hat{y} = \frac{1}{1+e^{-(w_{36}\hat{x}_3 + w_{46}\hat{x}_4 + w_{56}\hat{x}_5 + b_6)}} = \frac{1}{1+e^{-((0.01)(0) + (0.05)(0) + (0.02)(0.057) - 0.01)}} = 0.4978$$

Thus, for Age = 0.375 and Balance = 0.00199, this neural network produces a predicted probability of 0.4978 that this customer defaults on their loan. Suppose that this customer actually defaulted on their loan. For a cutoff value of 0.5, the neural network with these initial neuron weights and bias terms would mis-classify this customer, but just barely. In the next section, we describe how the neural network iteratively updates the values of its neuron weights and bias terms based on the difference between its predicted probabilities and the actual observed classes in an effort to improve its predictions.

How a Neural Network Learns

To help explain the process of how the parameters of a neural network are estimated, consider a training set of 10 observations from the Optiva example shown in Table 11.11. The variable values for these 10 observations have been scaled according to equation (11.7).

The process of training a neural network begins with the initialization of values for neuron weights and bias terms. These initial parameters are typically set to small random values between −0.5 and 0.5 such as those shown in Figure 11.18.

A single iteration of the neural network training process consists of the following steps:

1. Predict the probability of Class 1 for each observation in a batch using the current parameter values.
2. Update the parameter values based on the prediction error on the current batch of training set observations.

Table 11.11 Training Set Observations for the Neural Network

Customer	Loan Default	Age (years)	Balance ($)	Age (scaled)	Balance (scaled)
1	1	38	49	0.37500	0.00199
2	1	26	671	0.00000	0.09712
3	1	47	772	0.65625	0.11257
4	1	48	136	0.68750	0.01530
5	1	40	123	0.43750	0.01331
6	0	29	36	0.09375	0.00000
7	0	31	192	0.15625	0.02386
8	0	35	6,574	0.28125	1.00000
9	0	58	2,200	1.00000	0.33099
10	0	30	2,100	0.12500	0.31569
	Max:	58	6,574		
	Min:	26	36		

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 11.12 Results of First Iteration of Neural Network Training

Customer	Age (scaled)	Balance (scaled)	Loan Default (y)	Predicted Probability (ŷ)	Log Loss
1	0.3750	0.0020	1	0.49779	0.6976
2	0.0938	0.0000	0	0.49776	0.6887
3	0.4375	0.0133	1	0.49779	0.6976
4	0.2813	1.0000	0	0.49773	0.6886
5	0.6875	0.0153	1	0.49782	0.6975
				Average:	0.6940

The **batch size** is the number of training set observations whose outcome values are predicted before updating the parameter values. The batch size is a user-defined setting and can be as small as one (corresponding to the case in which the parameter values are updated after predicting a single observation) to the size of the training set (corresponding to the case in which the parameter values are updated after predicting all the observations in the training set).

Now we demonstrate the neural network training process using the training set in Table 11.11 using a batch size of five. The first training iteration considers observations 1–5. Starting with the initial parameter values in Figure 11.18, Table 11.12 contains the predicted probabilities of loan default for the first iteration of five observations.

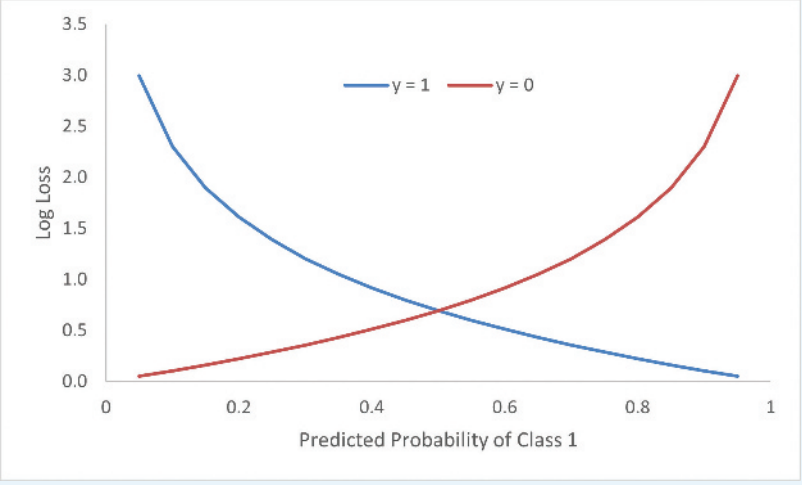
We measure the quality of the predicted probabilities in Table 11.12 using **log loss**. For a binary outcome y, the log loss of a predicted Class 1 probability $\hat{y}$ is computed as

$$\text{log loss} = -y \log \hat{y} - (1-y) \log (1-\hat{y})$$

As Figure 11.19 shows log loss values for a range of predicted Class 1 probabilities on both a Class1 observation ($y = 1$) and a Class 0 observation ($y = 0$). For a Class 1 observation, the value of log loss decreases as the predicted probability of Class 1 increases. For a Class 0 observation, the value of log loss increases as the predicted probability of Class 1 increases.

In Table 11.12, the quality of the predictions for observations 1–5 is measured by the average log loss. Recall that the neural network generates each predicted probability through

Figure 11.19 Log Loss for Range of Predicted Probabilities



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a forward pass of prediction calculations from the input layer to the output layer. In contrast, the parameters of the neural network (neuron weights and bias terms) are updated through a backward pass of adjustment calculations from the output layer to the input layer. A method called **backpropagation** employs the chain rule[4] from calculus to estimate how much a change in each of the respective parameters (neuron weights, bias terms) affects the prediction error of a batch of training set observations. Data mining software implements backpropagation to determine the amount to adjust each neuron weight and bias term. Let Δ_{ij} be the amount that the neuron weight w_{ij} is updated. Let Δ_i be the amount that the bias term b_i is updated. Then, the update formulas for a neuron weight w_{ij} and a bias term b_i are given by

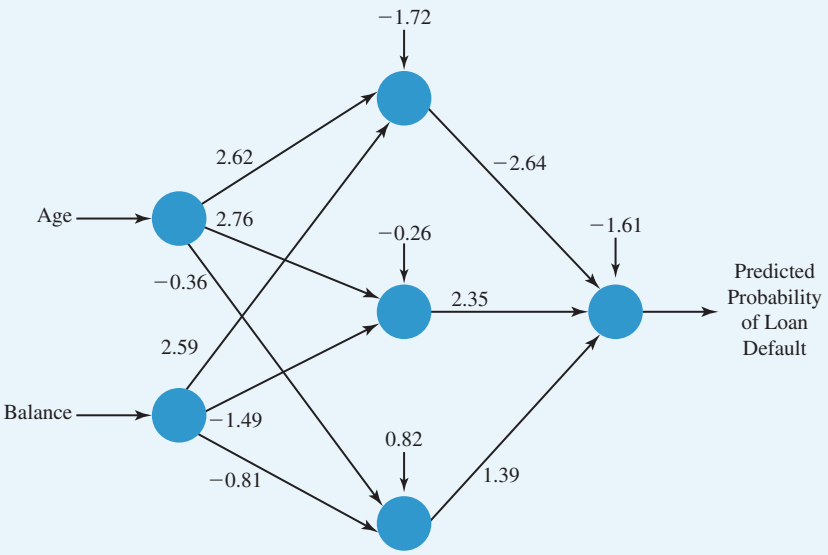
$$w_{ij}^{new} = w_{ij} + \Delta_{ij} \quad \textbf{(11.8)}$$

$$b_i^{new} = b_i + \Delta_i \quad \textbf{(11.9)}$$

After completing the update of all parameters, the next iteration begins by using the new values of the parameters to generate predictions for the next batch of observations (customers 6–10). Then, the log loss of the predicted probabilities on observations 6–10 is used to again adjust the parameters according to the update formulas in equations (11.8) and (11.9). These new values of the parameters are then used to generate predictions for the next batch of observations (customers 1–5 again). This process repeats until completing the specified number of iterations.

After several iterations, the neuron weights and bias terms in Figure 11.20 are obtained. The predicted probabilities in Table 11.13 are generated using the neuron weights and bias terms in Figure 11.20 using the ReLu activation function for the neurons in the hidden layer and the logistic activation function for the neuron in the output layer. By comparing the average log loss of 0.6940 from the first iteration of neural network training in Table 11.12 to the average log loss of 0.3401 after several iterations of neural network training, we observe that the neural network's predictions have improved. Based on a cutoff value of 0.5, the neural network would only misclassify one observation, customer 8.

Figure 11.20 Neural Network with Updated Neuron Weights and Bias Terms



[4] The chain rule is a formula for finding the derivative of the composition of two functions. The specifics of the chain rule, and its application to neural networks, are beyond the scope of this text. For more information see C. C. Aggarwal, *Neural Networks and Deep Learning*, Springer, 2018.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 11.13 Training Set Predictions by Neural Network with Updated Weights and Bias Terms

Customer	Age	Balance	Loan Default (y)	Predicted Probability ($\hat{y}$)	Log Loss
1	0.3750	0.0020	1	0.7586	0.2763
2	0.0938	0.0000	0	0.3712	0.4639
3	0.4375	0.0133	1	0.8125	0.2076
4	0.2813	1.0000	0	0.0029	0.0029
5	0.6875	0.0153	1	0.9335	0.0688
6	0.1563	0.0239	0	0.4344	0.5698
7	1.0000	0.3310	0	0.2194	0.2477
8	0.0000	0.0971	1	0.3567	1.0308
9	0.6563	0.1126	1	0.8256	0.1917
10	0.1250	0.3157	0	0.2892	0.3414
				Average:	0.3401

To evaluate the out-of-sample predictive accuracy of the neural network with the values of neuron weights and bias terms shown in Figure 11.20, we apply the neural network to the 10 validation set observations in Table 11.14. For these 10 observations, the average log loss is 0.4432. Based on a cutoff value of 0.5, the neural network would only misclassify one observation, customer 18.

Neural networks have the advantage of being able to capture complex relationships between the input features and the outcome variable. The disadvantage is that neural networks typically need more data (more observations) to effectively estimate the neuron weights and bias terms necessary to capture these complex relationships. Additionally, a neural network is not easily interpreted and often treated as a "black box" with model construction consisting of comprehensive testing of the many algorithm choices including the number of hidden layers, number of neurons in a hidden layer, activation functions used, etc.

Table 11.14 Validation Set Predictions by Neural Network with Updated Weights and Bias Terms

Customer	Age	Balance	Loan Default (y)	Predicted Probability ($\hat{y}$)	Log Loss
11	0.30435	0.04738	1	0.5337	0.4693
12	1.00000	0.14454	1	0.9904	0.3500
13	0.17391	0.00000	0	0.3472	0.6695
14	0.00000	0.32418	0	0.3567	0.3572
15	0.08696	0.40912	0	0.3329	0.3172
16	0.00000	0.32785	0	0.3566	0.3559
17	0.13044	1.00000	0	0.3226	0.0081
18	0.95652	0.18548	0	0.9524	0.9373
19	0.39130	0.27591	0	0.3226	0.6570
20	0.69565	0.15676	1	0.7479	0.3103
				Average:	0.4432

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notes + Comments

1. The training and validation sets used in this section to demonstrate neural networks are impractically small. We use these small data sets to simply illustrate the process.
2. In addition to the feed forward neural network architecture discussed in this section in which information flows unidirectionally, there is another type of structure called a recurrent neural network. A recurrent neural network allows information to flow in both directions by introducing loops between the neurons of the network.
3. In addition to the ReLu function, other common functions used as activation functions for the hidden layer of neural networks include the logistic function, $f(x) = \frac{1}{1+e^{-x}}$, the hyperbolic tangent function, $f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$, and the softplus function, $f(x) = \ln(1 + e^{x})$.
4. Scaling the input variables is an important pre-processing step in neural network classification because it facilitates the updating of the parameters as the neural network learns from the training data. Min–max scaling or standardizing a variable value by subtracting the mean and dividing by the standard deviation are common techniques. The best way to scale the variables may be dependent on the data, so experimenting with both ways is recommended.

11.7 Feature Selection

In this chapter, we have described how to apply logistic regression, k-nearest neighbors, classification trees, and neural networks to a classification task. These methods each generate a class prediction for a categorical outcome variable y by using the information in the set of input variables $x_1, x_2, \ldots, x_q$ in a different manner. In this section, we discuss how to select the input variables to include in the construction of a predictive model, a process called **feature selection**. Feature selection can be particularly important when dealing with data sets with many variables because including all the variables in the predictive model can lead to inferior predictions due to the presence of redundant and irrelevant variables.

There are three types of feature selection methods: wrapper methods, filter methods, and embedded methods. The choice of feature selection method depends on the type of predictive method and the data.

Wrapper Methods

Given data on a categorical outcome variable y and input variables $x_1, x_2, \ldots, x_m$, a **wrapper method** iteratively executes the following steps:

1. Select a subset of $q \leq m$ input variables.
2. Conduct the training and validation of a classification model considering the q selected input variables.

A wrapper method repeats steps 1 and 2 for a specified number of iterations. The simplest wrapper method considers all possible subsets of variables, which is called the **best subsets method**. The best subsets method is guaranteed to identify the subset of input variables which results in the best out-of-sample predictive performance. However, the best subsets approach requires 2^m iterations of training and validation, which is impractical for data sets with many variables.

Rather than exhaustively considering all possible subsets of variables, wrapper methods are designed to try to intelligently select the next subset of variables to consider at an iteration. Given enough time to evaluate different subsets of variables, a wrapper method often identifies a good (if not the best) subset of variables. However, wrapper methods may require a substantial amount of time to search among possible subsets of input variables when m is large, and unless the method considers all 2^m subsets, it cannot guarantee finding the set of input variables $x_1, x_2, \ldots, x_q$ that achieves the best out-of-sample predictive accuracy.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Filter Methods

Given data on an outcome variable y and input variables $x_1, x_2, \ldots, x_m$, a **filter method** uses statistical analysis to screen out the input variables that seem likely to be unhelpful in predicting the outcome variable. To identify variables that are potentially irrelevant, filter methods consider the statistical analysis of the relationship between the quantitative outcome variable y and an input variable x.

We discuss boxplots in Chapter 2.

If the input variable x under consideration is quantitative, then its relationship with the categorical outcome variable y can be visualized by constructing boxplots of x for each category of y. If the resulting boxplots for each category are very similar, then the input variable x may not add information and possibly can be removed. In addition, it is common to measure its relationship to the categorical outcome variable y by computing a **confidence interval** on the mean of x for each categorical value of y. A confidence interval on the mean is an interval estimate that is believed to contain the value of the population mean at some level of confidence. If the difference between the sample mean values of x for the different categories of y is not statistically significant, then the input variable x may be irrelevant and possibly removed. However, the level of statistical significance that reliably identifies an irrelevant input variable varies and is up to the analyst's discretion.

We discuss confidence intervals in Chapter 7.

We discuss hypothesis tests in Chapter 7.

If the input variable x under consideration is categorical, then it is common to measure its relationship to the categorical outcome variable y with a **test of independence**. A test of independence between two categorical variables x and y is a hypothesis test that evaluates if the difference in the number of observations with each combination of categorical values of x and y significantly differs from the number of observations one would expect if these variables were independent. If the difference between the sample frequencies and the expected frequencies is not statistically significant, then the input variable x may be irrelevant and possibly removed. However, the level of statistical significance that reliably identifies an irrelevant categorical input variable varies and is up to the analyst's discretion.

The advantage of filter methods is that they are relatively easy and fast to execute. A disadvantage is that the statistical measures employed in filter methods do not explicitly consider the performance of the predictive model and require the analyst to determine the thresholds for determining irrelevance. Therefore, filter methods may fail to identify the best collection of input variables.

Embedded Methods

Given data on an outcome variable y and input variables $x_1, x_2, \ldots, x_m$, an **embedded method** simultaneously selects the features to consider in the model as it trains the model. An embedded method is tailored to the predictive model that is being considered.

One example of a technique that can be used for embedded feature selection is regularization. **Regularization** is a general technique to avoid overfitting a predictive model by controlling the model complexity. In this section, we describe penalty-based regularization methods.

For example, in logistic regression model complexity is determined by the number of input variables. Recall that logistic regression uses the n observations of a training set to estimate coefficients $b_0, b_1, \ldots, b_q$ of q input variables to generate the predicted probability of an observation having outcome class 1 as computed by:

$$\text{Predicted class 1 probability} = \hat{y} = \frac{1}{1 - e^{-(b_0 + b_1 x_1 + \ldots b_q x_q)}}$$

Letting y_i be the actual class of observation i in the training set, the values of the coefficients $b_0, b_1, \ldots, b_q$ are estimated to minimize the log loss:

$$\sum_{i=1}^{n} -y_i \log(\hat{y}_i) - (1 - y_i)\log(1 - \hat{y}_i) \qquad \textbf{(11.10)}$$

A regularization approach to logistic regression is to add a penalty to equation (11.10) so that the coefficients $b_0, b_1, \ldots, b_q$ are estimated to minimize

$$\sum_{i=1}^{n} -y_i \log(\hat{y}_i) - (1 - y_i)\log(1 - \hat{y}_i) + \lambda \sum_{j=1}^{q} |b_j|^p \qquad \textbf{(11.11)}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In equation (11.11), λ is a tuning parameter and p is a parameter that is commonly set to either 1 or 2. For a value of $p = 1$, equation (11.11) is called **lasso regularization**. For a value of $p = 2$, equation (11.11) is called **ridge regularization**. Because the coefficients $b_0, b_1, \ldots, b_q$ are estimated to minimize equation (11.11), the addition of the regularization penalty tends to make at least some of the coefficient values closer to zero. The value of λ in equation (11.11) that results in the values of $b_0, b_1, \ldots, b_q$ that achieve the best out-of-sample predictive performance is determined by iteratively varying λ in a series of cross-validation experiments. For this best value of λ, the values of the coefficients $b_1, \ldots, b_q$ are examined. For logistic regression, lasso regularization serves as embedded feature selection as it often results in coefficient values equal to zero for one or more variables, suggesting that these variables may be dropped from the model.

Similar to how regularized logistic regression estimates its coefficients to minimize equation (11.11), regularization may be embedded in the training of a neural network. For a classification task, a regularized neural network estimates its parameters (neuron weights and bias terms) to minimize an equation analogous to equation (11.11), in which the regression coefficients are replaced with neural network parameters. Such a regularization approach helps avoid overfitting the neural network to the training data by forcing at least some of the weight values closer to zero.

Another example of embedded feature selection occurs in the construction of a classification tree. Recall that as a classification tree is being constructed on training data, the tree logic splits a node of observations by considering each input variable to identify the variable which results in the subsets with largest decrease in impurity of the outcome variable. The presence of extraneous input variables generally does not affect the construction of the classification tree. This is because an extraneous variable should not be selected as the basis of a splitting rule as it will not decrease the impurity of the outcome variable in the resulting subsets.

While feature selection is embedded in the construction logic of a classification tree, regularization can be useful to prune classification trees. In a classification tree, model complexity is determined by the number of rules. Recall that for a classification tree, the n observations in the training data are used to a set of rules that divide the observations into groups and then these rules are used to classify observations. A regularization approach to constructing a classification tree T is to determine the set of rules on the training set to minimize

$$R(T) + \lambda |T| \tag{11.12}$$

In equation (11.12), R(T) denotes the training set classification error of tree T, λ is a complexity parameter and $|T|$ corresponds to the number of leaf nodes in the tree. The regularization penalty in equation (11.12) makes trees with more branches more costly, so increasing the value of λ implicitly prunes the classification tree. On the other hand, if λ is set to zero, a full classification tree is constructed. The value of λ in equation (10.6) that results in the tree that achieves the best out-of-sample predictive accuracy is determined by iteratively varying λ in a series of cross-validation experiments.

Notes + Comments

1. Feature selection can be critical when performing k-nearest neighbors classification. When computing similarity on the basis of too many features, the distance between observations becomes less distinct with each additional feature and thus less helpful in making predictions.
2. Although classification trees have embedded feature selection, the appropriate representation of the variables (part of the data preparation phase) is important as regression trees are susceptible to poorly coded categorical variables, particularly categorical variables with many values.
3. Feature selection can improve neural network classification by reducing the number of parameters (feature weights and constant factors) that need to be estimated from the training data.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary

In this chapter, we introduced data mining concepts and techniques for classification tasks which involve a binary categorical variable. We began by describing how to partition data into training, validation, and test sets in order to construct and evaluate classification models. Specifically, we described both the static holdout set method and the *k*-fold cross-validation method. We also discussed the situation in which the two outcome classes are highly imbalanced and described ways to address this issue. We discussed various performance measures for classification and why measures other than accuracy may be preferred.

We presented several different predictive methods for a classification task. First, we explained how logistic regression is analogous to multiple linear regression for the case when the outcome variable is binary. Next, we described the classification of a binary categorical outcome using *k*-nearest neighbors. Then, we described classification trees and how combining many classification trees with an ensemble method generally outperforms predictive methods based on a single classifier. The last predictive method for a classification task we consider was neural networks. We showed how a neural network is a generalization of the logistic regression model and described the process of training a neural network and making predictions with it. We conclude the chapter with a discussion of methods for selecting features for predictive models and the use of regularization techniques to avoid overfitting.

Glossary

Accuracy Measure of classification success defined as 1 minus the overall error rate.
Activation function A function in a neuron of a neural network that produces an output from a neuron based on the inputs into a neuron.
Area under the ROC curve (AUC) A measure of a classification method's performance; an AUC of 0.5 implies that a method is no better than random classification while a perfect classifier has an AUC of 1.0.
Backpropagation A process by which the parameters of the neural network (neuron weights and bias terms) are updated through a backward pass of adjustment calculations from the output layer to the input layer.
Bagging An ensemble method that generates a committee of models each based on a slightly different training set and makes predictions based on the average prediction of the set of models.
Batch size The number of training set observations whose outcome values are predicted before updating the parameter values in a neural network.
Best subsets method A wrapper method for selecting input variables for a predictive model that considers all possible 2^m subsets of variables (where m is the total number of variables under consideration).
Bias term A parameter value that amplifies or dampens the input into a neuron of a neural network.
Boosting An ensemble method that iteratively adapts the construction of the *j*th training set based on the prediction error of the classifiers constructed on the previous $j-1$ training sets.
Class 0 error rate The percentage of Class 0 observations misclassified by a model in a data set.
Class 1 error rate The percentage of actual Class 1 observations misclassified by a model in a data set.
Classification A predictive data mining task requiring the prediction of an observation's outcome class or category.
Classification tree A tree that classifies a categorical outcome variable by splitting observations into groups via a sequence of hierarchical rules on the input variables.
Confidence interval An estimate of a population parameter (e.g., mean) that provides an interval believed to contain the value of the parameter at some level of confidence.
Confusion matrix A table summarizing the correct and incorrect predictions of a classification method.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Cutoff value The smallest value that the predicted probability of an observation can be for the observation to be classified as Class 1.
Decision boundary For a binary classification method, a boundary that separates the predictions of the two different classes.
Deep learning An artificial intelligence method based on specialized neural networks with many hidden layers used to discover intricate patterns in data.
Embedded method A method that selects features for a predictive model as part of the model training and validation process.
Ensemble method A predictive data mining approach in which a committee of individual classification models is generated and a prediction is made by combining these individual predictions.
F1 Score A measure combining precision and sensitivity into a single metric.
False negative The misclassification of a Class 1 observation as Class 0.
False positive The misclassification of a Class 0 observation as Class 1.
Feature An input variable used for predicting values of an outcome variable.
Feature selection The process of determining the input variables to include in the construction of a predictive model.
Feed forward neural network A neural network in which the prediction calculations are processed in only one direction.
Filter method A method for selecting features for a predictive model by analyzing the statistical relationships of the features and screening out variables likely to be irrelevant.
Hidden layer A layer of neurons in a neural network between the input layer and the output layer.
Impurity Measure of the heterogeneity of observations in a classification tree.
Input layer The first layer of neurons in a neural network through which the values of the input features enter.
***k*-fold cross-validation** A procedure to train and validate models in which the observations to be used to train and validate the model are repeatedly randomly divided into k subsets called folds. In each iteration, one fold is designated as the validation set and the remaining $k - 1$ folds are designated as the training set. The results of the iterations are then combined and evaluated.
***k*-nearest neighbors (*k*-NN)** A data mining method that predicts an observation i's outcome value based on the k observations most similar to observation i with respect to the input variables.
Lasso regularization Regularization in which the penalty term considers the absolute value function.
Lazy learner A predictive model that performs no calculations on the training set prior to being supplied an observation to predict.
Leaf node A node in a regression tree that is not further split into more nodes.
Leave-one-out cross-validation A special case of k-fold cross validation for which the number of folds equals the number of observations in the combined training and validation data.
Lift chart A chart used to compare how well a model performs in identifying Class 1 observations relative to randomly guessing.
Logistic function An activation function for a neural network of the following form $1/(1 + e^{-x})$
Logistic regression A generalization of linear regression that predicts a categorical outcome variable by computing the log odds of the outcome as a linear function of the input variables.
Log loss A measure indicating how close a classification model's prediction probability is to the corresponding actual value (with 0 representing the "no" class and 1 representing the "yes" class).
Min–max scaling The scaling of a variable so that its values lie in the range between 0 and 1.
Neural network A predictive data mining model whose structure is motivated by the biological functioning of the brain.
Neuron A node in a neural network.
Neuron weight A parameter value representing the strength of a neuron's output signal to a neuron in the next layer of a neural network.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Out-of-bag estimation A measure of estimating the predictive performance of a bagging ensemble of m models (without a separate validation set) by leveraging the concept that the training of each model is only based on approximately 63.2% of the original observations (due to sampling with replacement).
Output layer The final layer of neurons in a neural network through which the predicted outcome values are obtained.
Overall error rate The percentage of observations misclassified by a model in a data set.
Overfitting A situation in which a model explains random patterns in the data on which it is trained rather than just the generalizable relationships, resulting in a model with training-set performance that greatly exceeds its performance on out-of-sample data.
Oversampling A technique that balances the number of Class 1 and Class 0 observations in a training set by inserting copies of minority class observations into the training set.
Perceptron A neural network consisting only of an input layer and output layer (no hidden layers).
Precision The percentage of observations predicted to be Class 1 that actually are Class 1.
Pruning The process of obtaining a tree with fewer leaf nodes than a full tree to prevent overfitting on training data.
Random forest A variant of the bagging ensemble method that generates a committee of trees based on different random samples but restricts each individual tree to a limited number of randomly selected features (variables).
Receiver operating characteristic (ROC) curve A chart used to illustrate the tradeoff between a model's ability to identify Class 1 observations and its Class 0 error rate.
Rectified linear unit (ReLu) function A function $f(x)$ that is equal to 0 if $x < 0$ and equal to x if $x > 0$.
Regularization A penalty-based method to avoid overfitting a predictive model by controlling model complexity.
Ridge regularization Regularization in which the penalty term considers a quadratic function.
Root node The top node in a classification tree containing all the observations in the training set.
Sensitivity (recall) The percentage of actual Class 1 observations correctly identified.
Specificity The percentage of actual Class 0 observations correctly identified.
Stratified random sampling A data sampling method that ensures that resulting samples (for training, validation, and test sets) contain the same outcome class distribution as the source data.
Stump A tree with a only a single rule.
Supervised learning Category of data mining techniques in which an algorithm learns how to predict an outcome variable of interest.
Test of independence A test between two categorical variables and y that evaluates if the difference in the number of observations with each combination of categorical values of and y significantly differs from the number of observations one would expect if these variables were independent.
Test set Data set used to compute unbiased estimate of final predictive model's performance.
Training set Data used to construct candidate predictive models.
Undersampling A technique that balances the number of Class 1 and Class 0 observations in a training set by removing majority class observations from the training set.
Unstable learner A prediction model for which small changes in the training set cause a model's predictions to fluctuate substantially.
Validation set Data used to evaluate candidate predictive models.
Weak learner A predictive model whose individual predictions are only slightly better than random guessing.
Wrapper method A method for selecting features for a predictive model that iteratively tests different subsets of features.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Problems

Problems 1 through 29 do not require the use of data mining software and focus on knowledge of concepts and basic calculations.

Conceptual

1. **Classification with Logistic Regression.** Which of the following statements about logistic regression is true? LO 3
   a. Logistic regression estimates the probability of an observation being in the "yes" class as a linear function of input variables.
   b. Logistic regression does not assume a functional relationship between a binary event and the input variables.
   c. Logistic regression generates a linear decision boundary to classify observations.
   d. If the p-value of an input variable in a logistic regression model exceeds 0.05, then the input variable is unhelpful in classifying observations and should be removed from the model.

2. **Logistic Regression Decision Boundary.** In Section 11.3, the logistic regression equation to predict whether a movie wins the Best Picture Academy Award based on the total number of Oscar nominations was:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -6.214 + 0.596 \text{ Oscar Nominations}$$

   For a cutoff value of $\hat{p} = 0.5$, what is the decision boundary separating Class 1 from Class 0 observations for this model? Demonstrate that this agrees with the classifications in Table 11.4.

3. **Classification with *k*-Nearest Neighbors.** Which of the following statements about the k-nearest neighbors classifier is true? LO 4
   a. The decision boundary corresponding to a k-nearest neighbor classifier is a linear function of the input variables.
   b. The decision boundary corresponding to a k-nearest neighbor classifier can be a highly nonlinear function of the input variables.
   c. When applying k-nearest neighbors, one should include as many input variables as possible.
   d. Standardizing the input variables is not necessary before applying a k-nearest neighbor classifier.

4. **Accuracy Measure.** For 10,000 observations, suppose a classifier generates 235 false positives, 138 false negatives, 9,432 true negatives, and 195 true positives. What is the classifier's accuracy? LO 2

5. **Sensitivity Measure.** For 10,000 observations, suppose a classifier generates 235 false positives, 138 false negatives, 9,432 true negatives, and 195 true positives. What is the classifier's sensitivity? LO 2

6. **Specificity Measure.** For 10,000 observations, suppose a classifier generates 235 false positives, 138 false negatives, 9,432 true negatives, and 195 true positives. What is the classifier's specificity? LO 2

7. **Precision Measure.** For 10,000 observations, suppose a classifier generates 235 false positives, 138 false negatives, 9,432 true negatives, and 195 true positives.What is the classifier's precision? LO 2

8. **Performance Measures.** An analyst for an insurance company's data science team is developing a model to predict whether a submitted claim is fraudulent. Generally, less than 10% of claims are fraudulent. The threshold value for classifying an observation as fraudulent is not universally agreed upon. Which of the following is a good way to evaluate a model for classifying claims as fraudulent or not? LO 2
   a. Precision
   b. Specificity
   c. Accuracy
   d. Area under the ROC curve

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

9. **Effect of Increasing Classification Threshold.** When the threshold value for binary classification increases, what happens? LO 2
   a. Sensitivity decreases and specificity increases.
   b. F1 score increases.
   c. Precision increases.
   d. Accuracy decreases.
10. **Lift Measure.** A sports betting company is testing the effectiveness of its online ads and has developed a classification model that predicts whether a potential customer will respond to its offer using information in the potential customer's browser history (cookies, etc.). Out of a test dataset of 1,000 customers, the 100 customers most likely to respond (according to the model) are identified and it is discovered that 11 of these actually did respond to the online ads. In the entire test dataset of 1,000 customers, only 30 customers responded to the ad. What is the lift of the classification model for the set of 100 customers it identified as most likely to respond? LO 2
11. **Interpreting Lift.** A sports betting company is testing the effectiveness of its online ads and has developed a classification model predicts whether a potential customer will respond to its offer using information in the potential customer's browser history (cookies, etc.). For the top 10% of customers most likely to respond to online ads in a test dataset, an analyst determines that the lift of the classification model is 5.75. What is a correct interpretation of lift for this subset of customers? LO 2
    a. The classification model identified 5.75 times more customers that responded to online ads than if a random 10% of customers was selected.
    b. The classification model identified 575% more customers that responded to online ads than if a random 10% of customers was selected.
    c. The classification model identified 5.75% more customers that responded to online ads than if a random 10% of customers was selected.
    d. Over the entire range of threshold values, the classification model is 5.75 more likely to identify a responding customer than a non-responding customer.
12. **Full Classification Tree.** What happens if a classification tree is grown to its fullest extent? LO 5
    a. Each leaf node will be 100% pure (all Class 1 observations or all Class 0 observations)
    b. The best predictive performance on out-of-sample data will be achieved.
    c. Each leaf node will consist of a single observation.
    d. Training error is minimized, but the tree is very likely to be overfit.
13. **Instability of Classification Trees.** It is known that tree-based learning is unstable. What does this mean? LO 5
    a. Trees are poor predictive models.
    b. The predictive performance of the tree varies dramatically based on the training data used.
    c. When the data used to train a tree is changed (even slightly), the variables (and their split values) composing the tree structure may vary dramatically.
    d. Trees are difficult to interpret and explain.
14. **Feature Selection in Classification Trees.** How are variables selected in tree-based learning? LO 5
    a. Variables are automatically selected by the tree logic as the best variable (and value to split it at) is identified at each step of the recursive partitioning.
    b. Variables must be selected in a preprocessing step using a filter method.
    c. A regularization penalty is factored in the impurity measure of each variable split during the induction process.
    d. Cross-validation is used to select the variables used in a tree.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

15. **Bagging Ensembles.** Which of the following statements about bagging is false? LO 6
    a. Bagging generates a composite prediction by weighting the votes of the individual classifiers by their individual accuracies.
    b. Bagging generates a committee of different classifiers by constructing them from slightly different training sets.
    c. Bagging enables out-of-bag error estimation that provides an estimate of out-of-sample predictive performance.
    d. A weakness of bagging is that the individual models still may be highly correlated.
16. **Boosting Ensembles.** Which of the following statements about boosting is false? LO 6
    a. Boosting works best when using logistic regression models as the individual classifiers.
    b. Boosting constructs a sequence of classifiers for which the training set for classifier $j + 1$ depends on the prediction errors of classifiers j on its training set.
    c. Boosting predictions are based on the votes of individual classifiers in which each vote is weighted proportional to the corresponding classifier's training set accuracy.
    d. The number of classifiers in a boosting committee is a parameter that must be determined via cross-validation to prevent overfitting.
17. **Ensemble Accuracy. Ensembles.** Suppose each individual classifier in an ensemble has a 70% chance of predicting an observation's class correctly and that the individual classifiers are perfectly independent of each other. What is the probability that an ensemble of 11 such classifiers correctly predicts an observation's class? LO 6
    a. 70%
    b. 92.2%
    c. 57.2%
    d. 11.8%
18. **Individual Classifier Type in an Ensemble.** Why are trees often used as individual classifiers in an ensemble approach? LO 6
    a. Trees are computationally easy to construct.
    b. Trees are easy to interpret.
    c. Trees have embedded feature selection.
    d. Trees are unstable predictors.
19. **Random Forests.** Which of the following statements about random forests is false? LO 6
    a. The predictive performance of a random forest is often competitive with a boosting ensemble, and it is computationally much simpler.
    b. A random forest is an ensemble consisting of a combination of different types of classifiers (k-nearest neighbors, logistic regression, neural networks, and trees).
    c. A random forest is not as interpretable as an individual tree.
    d. Correlation between individual trees in a random forest is reduced by limiting the number features considered at the variable splits.
20. **Neural Network Architecture.** Which of the following statements about the architecture of a neural network is false?
    a. The output layer of a neural network for a binary classification task produces an observation's probability of a Class 1 outcome.
    b. The activation function in the output layer of a neural network for a classification task is commonly the logistic function.
    c. The values of the input features do not need to be scaled or standardized.
    d. The number of neurons in the input layer typically corresponds to the number of input features being considered.
21. **Activation Functions.** Which function is a common choice for the output layer in a neural network for a classification task? LO 7
    a. Logistic function
    b. Identity function
    c. ReLu function
    d. Quadratic function

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

22. **Feature Selection.** In a filter method, which of the following is helpful to assess the relationship between a quantitative input variable and a categorical outcome variable? LO 8
    a. Test of independence
    b. ROC curve
    c. Scatter chart
    d. Boxplots

23. **Regularization.** For which classification method is a lasso regularization commonly used to select features? LO 9
    a. *k*-nearest neighbors
    b. Classification trees
    c. Logistic regression
    d. Random forest

24. **Dating Web Site.** The dating web site Oollama.com requires its users to create profiles based on a survey in which they rate their interest (on a scale from 0 to 3) in five categories: physical fitness, music, spirituality, education, and alcohol consumption. A new Oollama customer, Erin O'Shaughnessy, has reviewed the profiles of 100 prospective dates to indicate whether there was any interest in learning more about them.

    Oollama has split the 100 profiles rated "interested" or "not interested" by Erin into a training set of 60 observations and a validation set of 40 observations. Oollama then trained a logistic regression to predict Erin's interest in other profiles that Erin has not yet viewed. The resulting logistic regression model is as follows:

$$\text{Log Odds of Interested} = -0.920 + 0.325\ \text{Fitness} - 3.611\ \text{Music} + 5.535\ \text{Education} - 2.927\ \text{Alcohol}$$

    For 40 profiles (observations) in the validation set, this logistic regression model generates that following probability of Interested. LO 2, 3

Observation	Interested	Probability of Interested	Observation	Interested	Probability of Interested
35	1	1.000	13	0	0.412
21	1	0.999	2	0	0.285
29	1	0.999	3	0	0.219
25	1	0.999	7	0	0.168
39	1	0.999	9	0	0.168
26	1	0.990	12	0	0.168
23	1	0.981	18	0	0.168
33	1	0.974	22	1	0.168
1	0	0.882	31	1	0.168
24	1	0.882	6	0	0.128
28	1	0.882	20	0	0.128
36	1	0.882	15	0	0.029
16	0	0.791	5	0	0.020
27	1	0.791	14	0	0.015
30	1	0.791	19	0	0.011
32	1	0.791	8	0	0.008
34	1	0.791	10	0	0.001
37	1	0.791	17	0	0.001
40	1	0.791	4	0	0.001
38	1	0.732	11	0	0.000

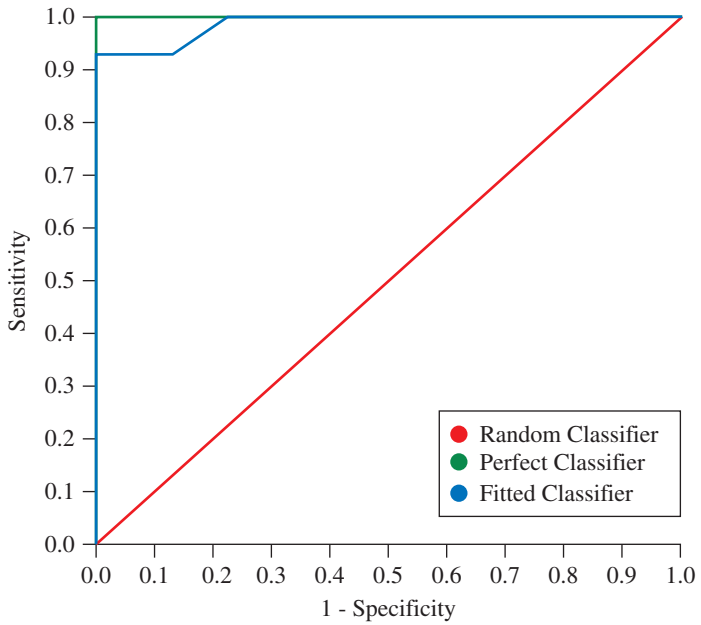
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Using a cutoff value of 0.5 to classify a profile observation as Interested or not, construct the confusion matrix for this 40-observation validation set. Compute sensitivity, specificity, and precision measures and interpret them within the context of Erin's dating prospects.
b. Oollama understands that its clients have a limited amount of time for dating and therefore use lift charts to evaluate their classification models. For these data, what is the lift resulting from the logistic regression model on the top 25% of profiles that Erin is most likely to be interested in? Interpret this value.
c. A recently posted profile has values of Fitness = 3, Music = 1, Education = 3, and Alcohol = 1. Use the estimated logistic regression equation to compute the probability of Erin's interest in this profile.

25. **Cupcake Approval (*k*-NN classification).** Fleur-de-Lis is a boutique bakery specializing in cupcakes. The bakers at Fleur-de-Lis like to experiment with different combinations of four major ingredients in its cupcakes and collect customer feedback; it has data on 150 combinations of ingredients with the corresponding customer reception for each combination classified as "thumbs up" (Class 1) or "thumbs down" (Class 0). To better anticipate the customer feedback of new recipes, Fleur-de-Lis has determined that a *k*-nearest neighbors classifier with $k = 10$ seems to perform well.

Using a cutoff value of 0.5 and a validation set of 45 observations, Fleur-de-Lis constructs following confusion matrix and the ROC curve for the *k*-nearest neighbors classifier with $k = 10$. **LO 2, 4**

	Predicted Feedback	
Actual Feedback	**Thumbs Up**	**Thumbs Down**
Thumbs Up	13	1
Thumbs Down	1	30



As the confusion matrix shows, there is one observation that actually received thumbs down, but the *k*-nearest neighbors classifier predicts a thumbs up. Also, there is

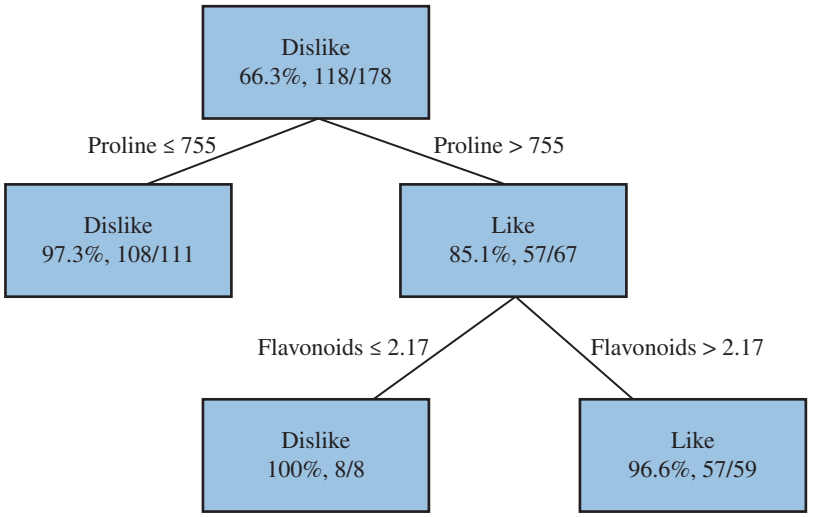
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

one observation that actually received a thumbs up, but the *k*-nearest neighbors classifier predicts a thumbs down. Specifically:

Observation ID	Actual Class	Probability of Thumbs Up	Predicted Class
A	Thumbs Down	0.5	Thumbs Up
B	Thumbs Up	0.2	Thumbs Down

a. Explain how the probability of Thumbs Up was computed for Observation A and Observation B. Why was Observation A classified as Thumbs Up and Observation B was classified as Thumbs Down?
b. Compute the values of sensitivity and specificity corresponding to the confusion matrix created using the cutoff value of 0.5. Locate the point corresponding to these values of sensitivity and specificity on the Fleur-de-Lis's ROC curve.
c. Based on what we know about Observation B, if the cutoff value is lowered to 0.2, what happens to the values of sensitivity and specificity? Explain. Use the ROC curve to estimate the values of sensitivity and specificity for a cutoff value of 0.2.

26. **Wine Approval.** Sommelier4U is a company that ships its customers bottles of different types of wine and then has the customers rate the wines as "Like" or "Dislike." For each customer, Sommelier4U trains a classification tree based on the characteristics of the wine such as amount of proline and flavonoids in the wine and customer ratings of wines that the customer has tasted. Then, Sommelier4U uses the classification tree to identify new wines that the customer may Like. Sommelier4U recommends the wines that have a greater than 50% probability of being liked. Neal Jones, a loyal customer, has provided "Like" or "Dislike" ratings on 178 different wines. Based on these 178 wines, Sommelier4U executed a 10-fold cross-validation experiment to obtain the following pruned classification tree. **LO 2, 5**



a. Construct the confusion matrix resulting from the pruned classification tree on the 178 wines.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Sommelier4U plans to apply the pruned classification tree and recommend wine that the tree believes that Neal will like. If Sommelier4U is most concerned that Neal will like the wines it recommends, which performance measure should it focus on? Calculate the value of this measure from the confusion matrix in part a.
c. Consider the wine with the following characteristics: Proline = 820 and Flavonoids = 3.1. Does Sommelier4U believe that Neal will like this wine?

27. **Alumni Donors.** A university is applying classification methods in order to identify alumni who may be interested in donating money. The university has a database of 58,205 alumni profiles containing numerous variables. Of these 58,205 alumni, only 576 have donated in the past. The university has oversampled the data and trained a random forest of 100 classification trees. For a cutoff value of 0.5, the following confusion matrix summarizes the performance of the random forest on a validation set: **LO 2, 6**

		Predicted
Actual	**Donation**	**No Donation**
Donation	268	20
No Donation	5375	23,439

The following table lists some information on individual observations from the validation set.

Observation ID	Actual Class	Probability of Donation	Predicted Class
A	Donation	0.8	Donation
B	No Donation	0.1	No Donation
C	No Donation	0.6	Donation

a. Explain how the Probability of Donation was computed for the three observations. Why were Observations A and C classified as Donation and Observation B was classified as No Donation?
b. Compute the values of accuracy, sensitivity, specificity, and precision. Explain why accuracy is a misleading measure to consider in this case. Evaluate the performance of the random forest, particularly commenting on the precision measure.

28. **Targeted Coupon Offers.** Honey is a technology company that provides online coupons to its subscribers. Honey's analytics staff has developed a classification method to predict whether a customer who has been sent a coupon will apply the coupon toward a purchase. For a sample of customers, the following table lists the classification model's estimated coupon usage probability for a customer. For this particular campaign, suppose that when a customer uses a coupon, Honey receives $0.20 in revenue from the product sponsor. To target the customer with the coupon offer, Honey incurs a cost of $0.01. Honey will offer a customer a coupon as long as the expected profit of doing so is positive. Use the following equation to determine which customers should be sent the coupon. **LO 2**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Expected Profit of Coupon Offer
$= P(\text{coupon used}) \times \text{Profit if coupon used}$
$+ (1 - P(\text{coupon used})) \times \text{Profit if coupon not used}$

Customer	Probability of Using Coupon
1	0.49
2	0.36
3	0.27
4	0.11
5	0.04

29. **Addressing Customer Churn.** Watershed is a media services company that provides online streaming movie and television content. As a result of the competitive market of streaming service providers, Watershed is interested in proactively identifying who will unsubscribe in the next three months based on the customer's characteristics. For a test set of customers, the file *watershed* contains an indication of whether a customer unsubscribed in the past three months and the classification model's estimated unsubscribe probability for the customer. In an effort to prevent customer churn, Watershed wishes to offer promotions to customers who may unsubscribe. It costs Watershed \$10 to offer a promotion to a customer. If offered a promotion, it successfully persuades a customer to remain a Watershed customer with probability 0.6, and retaining the customer is worth \$60 to Watershed. Customers will be offered the promotion in order of decreasing estimated unsubscribe probability. Compute the expected profit from offering the top n customers a promotion as:

DATA*file* watershed

$$\begin{aligned}\text{Profit} = {} & \text{Number of unsubscribing customers in top } n \\ & \times (P(\text{unsubscribing customer persuaded to remain}) \times (60 - 10) \\ & + P(\text{unsubscribing customer is not persuaded}) \times (0 - 10)) \\ & + \text{Number of customers who don't intend to unsubscribe} \times (0 - 10)\end{aligned}$$

Determine how many customers to which Watershed should offer the promotion to maximize the profit of the intervention campaign. **LO 2**

Software Application

Problems 30 through 56 and the Case Problem require the use of data mining software (such as R or Orange). Guidelines for solving these problems with R and Orange are provided in the respective appendixes available in the Cengage eBook.

30. **Coupon Use: Logistic Regression.** Salmons Stores operates a national chain of women's apparel stores. Five thousand copies of an expensive four-color sales catalog have been printed, and each catalog includes a coupon that provides a \$50 discount on purchases of \$200 or more. Salmons would like to send the catalogs only to customers who have the highest probability of using the coupon. For each of 1,000 Salmons customers, three variables were tracked from an earlier promotional campaign: last year's total spending at Salmons (Spending), whether they have a Salmons store credit card (Card), and whether they used the promotional coupon they were sent (Coupon). Apply logistic regression with lasso regularization to classify observations as a promotion-responder or not by using Coupon as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 3, 8, 9**

DATA*file* salmons.csv

a. Determine the lasso regularization penalty that maximizes AUC in a validation procedure.
b. For the level of lasso regularization identified in part (a), what are the values of the intercept and variable coefficients in this final model? Interpret the coefficients. Does the lasso regularization eliminate any variables?
c. For the final model selected in part (d), compute and interpret the lift measure on the top 10% of the test set observations most likely to use the promotional coupon.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
salmons.csv

31. **Coupon Use: *k*-Nearest Neighbors.** Refer to the scenario in Problem 30 regarding the identification of coupon promotion responders. Apply *k*-nearest neighbors to classify observations as a promotion-responder or not by using Coupon as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 4**
    a. Determine the value of *k* that maximizes the AUC in a validation procedure.
    b. For the best-performing *k*-nearest neighbors model in the validation procedure, compute and interpret the lift measure on the top 10% of the test set observations most likely to use the promotional coupon.

DATA *file*
salmons.csv

32. **Coupon Use: Classification Tree.** Refer to the scenario in Problem 30 regarding the identification of coupon promotion responders. Apply a classification tree to classify observations as a promotion-responder or not by using Coupon as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 5**
    a. Train a full classification tree and report its AUC from a validation experiment.
    b. Train a pruned classification tree and report its AUC from a validation experiment.
    c. Compare the AUC from the full classification tree in part (a) to the AUC from the pruned classification tree in part (b). Explain the difference.
    d. Consider a customer who spent $2,000 last year at Salmons and has a Salmons credit card. Using cutoff value of 0.50 with the pruned tree from part (b), what is the predicted credit score for this individual?

DATA *file*
salmons.csv

33. **Coupon Use: Neural Network.** Refer to the scenario in Problem 30 regarding the identification of coupon promotion responders. Apply a neural network to classify observations as a promotion-responder or not by using Coupon as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 7**
    a. Determine the neural network configuration that maximizes the AUC in a validation procedure.
    b. For the best-performing neural network identified in part (a), compute and interpret the lift measure on the top 10% of the test set observations most likely to use the promotional coupon.

DATA *file*
dana.csv

34. **Student Retention: Logistic Regression.** Over the past few years the percentage of students who leave Dana College at the end of their first year has increased. Last year, Dana started voluntary one-credit hour-long seminars with faculty to help first-year students establish an on-campus connection. If Dana is able to show that the seminars have a positive effect on retention, college administrators will be convinced to continue funding this initiative. Dana's administration also suspects that first-year students with lower high school GPAs have a higher probability of leaving Dana at the end of the first year. Data on the 500 first-year students from last year has been collected. Each observation consists of a first-year student's high school GPA, whether they enrolled in a seminar, and whether they dropped out and did not return to Dana. Apply logistic regression with lasso regularization to classify observations as dropped out or not by using Dropped as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 3, 8, 9**
    a. Determine the lasso regularization penalty that maximizes AUC in a validation procedure.
    b. For the level of lasso regularization identified in part (a), what are the values of the intercept and variable coefficients in this final model? Interpret the coefficients. Does the lasso regularization eliminate any variables?
    c. What does the final model selected in part (b) imply about the effect of the first-year seminars on retention?
    d. The data analyst team realized that they jumped directly into building a predictive model without exploring the data. Using descriptive statistics and charts, investigate

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the relationship between GPA and Seminar. For next year's first-year class, what could Dana's administration do regarding the enrollment of the seminars to better determine whether they have an effect on retention?

35. **Student Retention: *k*-Nearest Neighbors.** Refer to the scenario in Problem 34 regarding the identification of students who drop out of school. Apply *k*-nearest neighbors to classify observations as a drop-out or not by using Dropped as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 4**

DATA*file*
dana.csv

a. Determine the value of *k* that maximizes the AUC in a validation procedure.
b. For the best-performing *k*-nearest neighbors model identified in part (a), what is the sensitivity (recall) from the validation procedure. Interpret this measure.

36. **Student Retention: Classification Tree.** Refer to the scenario in Problem 34 regarding the identification of students who drop out of school. Apply a classification tree to classify observations as a drop-out or not by using Dropped as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 5**

DATA*file*
dana.csv

a. Train a full classification tree and report its AUC from a validation experiment.
b. Train a pruned classification tree and report its AUC from a validation experiment.
c. Compare the AUC from the full classification tree in part (a) to the AUC from the pruned classification tree in part (b). Explain the difference.

37. **Student Retention: Neural Network.** Refer to the scenario in Problem 34 regarding the identification of students who drop out of school. Apply a neural network to classify observations as a drop-out or not by using Dropped as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 7**

DATA*file*
dana.csv

a. Determine the neural network configuration that maximizes the AUC in a validation procedure.
b. For the best-performing neural network identified in part (a), what is the sensitivity (recall) from the validation procedure. Interpret this measure.

38. **Paperless Banking Adoption: Logistic Regression.** Sandhills Bank would like to increase the number of customers who use paperless banking as part of the rollout of its e-banking platform. Management has proposed offering an increased interest rate on a savings account if customers enroll for comprehensive paperless banking. To forecast the success of this proposal, management would like to estimate how many of the 200 current customers who do not use paperless banking would accept the offer. The IT company that handles Sandhills Bank's e-banking has provided anonymized data for 1,000 customers from one of its other client banks, Plains Bank, that made a similar promotion to increase paperless banking participation. For these 1,000 customers in the file *plains*, each observation consists of the average monthly checking account balance, the age of the customer, and whether the customer enrolled for paperless banking. Apply logistic regression with lasso regularization to classify observations in the file *plains* as enrolling in paperless banking or not by using Enroll as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 3, 8, 9**

DATA*file*
plains.csv
sandhills.csv

a. Determine the lasso regularization penalty that maximizes AUC in a validation procedure.
b. For the level of lasso regularization identified in part (a), what are the values of the intercept and variable coefficients in this final model? Interpret the coefficients. Does the lasso regularization eliminate any variables?
c. The file *sandhills* contains data corresponding to Sandhills Bank's 200 current customers. As Sandhills has not yet launched its promotion to any of these 200 customers, these observations only have data for the average monthly checking account balance and age of the customer, but no column for whether the customer

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

enrolled for paperless banking. Sandhills would like to estimate the likelihood of these customers enrolling for paperless banking. Apply the final model selected in part (e) to the 200 observations in the file *sandhills*. Using a cutoff value of 0.5, how many of Sandhills Bank's 200 customers does the final model predict will enroll for paperless banking?

39. **Paperless Banking Adoption: *k*-Nearest Neighbors.** Refer to the scenario in Problem 38 regarding the identification of bank customers who enroll in paperless banking. Apply *k*-nearest neighbors to classify observations as enrolling or not by using Enroll as the target (or response) variable. Use 100% of the data in the file *plains* for training and validation (do not use any data as a test set). **LO 1, 2, 4**

DATA *file*
plains.csv
sandhills.csv

a. Determine the value of *k* that maximizes the AUC in a validation procedure.
b. The file *sandhills* contains data corresponding to Sandhills Bank's 200 current customers. As Sandhills has not yet launched its promotion to any of these 200 customers, these observations only have data for the average monthly checking account balance and age of the customer, but no column for whether the customer enrolled for paperless banking. Sandhills would like to estimate the likelihood of these customers enrolling for paperless banking. Apply the final model selected in part (a) to the 200 observations in the file *sandhills*. Using a cutoff value of 0.5, how many of Sandhills Bank's 200 customers does the final model predict will enroll for paperless banking?

40. **Paperless Banking Adoption: Classification Tree.** Refer to the scenario in Problem 38 regarding the identification of bank customers who enroll in paperless banking. Apply a classification tree to classify observations as enroll or not by using Enroll as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 5**

DATA *file*
plains.csv
sandhills.csv

a. Train a pruned classification tree and report its AUC from a validation experiment.
b. The file *sandhills* contains data corresponding to Sandhills Bank's 200 current customers. As Sandhills has not yet launched its promotion to any of these 200 customers, these observations only have data for the average monthly checking account balance and age of the customer, but no column for whether the customer enrolled for paperless banking. Sandhills would like to estimate the likelihood of these customers enrolling for paperless banking. Apply the final model selected in part (a) to the 200 observations in the file *sandhills*. Using a cutoff value of 0.5, how many of Sandhills Bank's 200 customers does the final model predict will enroll for paperless banking?

41. **Paperless Banking Adoption: Neural Network.** Refer to the scenario in Problem 38 regarding the identification of bank customers who enroll in paperless banking. Apply a neural network to classify observations as a enroll or not by using Enroll as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 7**

DATA *file*
plains.csv
sandhills.csv

a. Determine the neural network configuration that maximizes the AUC in a validation procedure.
b. The file *sandhills* contains data corresponding to Sandhills Bank's 200 current customers. As Sandhills has not yet launched its promotion to any of these 200 customers, these observations only have data for the average monthly checking account balance and age of the customer, but no column for whether the customer enrolled for paperless banking. Sandhills would like to estimate the likelihood of these customers enrolling for paperless banking. Apply the final model selected in part (a) to the 200 observations in the file *sandhills*. Using a cutoff value of 0.5, how many of Sandhills Bank's 200 customers does the final model predict will enroll for paperless banking?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

42. **Undecided Voters: Logistic Regression.** Campaign organizers for both the Republican and Democratic parties are interested in identifying individual undecided voters who would consider voting for their party in an upcoming election. A non-partisan group has collected data on a sample of voters with tracked variables. The variables in this data are listed in the following table.

DATA*file*
blueorred.csv

Variable	Description
Vote	Voter status: Undecided or decided
Age	Age of voter
HomeOwner	Binary variable indicating whether voter owns home
Gender	Gender identity provided by voter
Married	Marital status of voter
HouseholdSize	Number of people in voter's household
Income	Annual income (in $1000)
Education	Years of education
Religious	Binary variable indicating whether voter regularly attends religious services

Apply logistic regression with lasso regularization to classify observations as being undecided or not by using Vote as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 3, 8,9**

a. Determine the lasso regularization penalty that maximizes AUC in a validation procedure.
b. Apply the final model selected in part (a) to the test set. What are the values of the intercept and variable coefficients in this final model? Interpret the coefficients.
c. For the final model selected in part (a), compute and interpret the precision on the test set.

43. **Undecided Voters: *k*-Nearest Neighbors.** Refer to the scenario in Problem 42 regarding the identification of undecided voters. Apply *k*-nearest neighbors to classify observations as undecided or not by using Vote as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 4, 8**

DATA*file*
blueorred.csv

a. Based on all the input variables, determine the value of *k* that maximizes the AUC in a validation procedure.
b. For the best-performing *k*-nearest neighbors model in the validation procedure (with respect to AUC), compute and interpret the precision on the test set.

44. **Undecided Voters: Classification Tree.** Refer to the scenario in Problem 42 regarding the identification of undecided voters. Apply a classification tree to classify observations as undecided or not by using Vote as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 5**

DATA*file*
blueorred.csv

a. Train a full classification tree and report its AUC from a validation experiment.
b. Train a pruned classification tree and report its AUC from a validation experiment.
c. Compare the AUC from the full classification tree in part (a) to the AUC from the pruned classification tree in part (b). Explain the difference.
d. Consider a 50-year-old man who attends church, has 15 years of education, owns a home, is married, lives in a household of four people, and has an annual income of

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$150,000. For a cutoff value of 0.5, does the pruned tree classify this observation as Undecided?

e. For the pruned tree, compute and interpret the precision on the test set.

45. **Undecided Voters: Random Forest.** Refer to the scenario in Problem 42 regarding the identification of undecided voters. Apply a random forest to classify observations as undecided or not by using Vote as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 6**

DATA*file* blueorred.csv

a. Determine the random forest configuration that maximizes the AUC in a validation procedure.

b. For the best-performing random forest identified in part (a), compute and interpret the precision on the test set.

46. **Undecided Voters: Neural Network.** Refer to the scenario in Problem 42 regarding the identification of undecided voters. Apply a neural network to classify observations as undecided or not by using Vote as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 7**

DATA*file* blueorred.csv

a. Determine the neural network configuration that maximizes the AUC in a validation procedure.

b. For the best-performing neural network identified in part (a), compute and interpret the precision on the test set.

47. **Cellphone Customer Retention: Logistic Regression.** Telecommunications companies providing cell-phone service are interested in customer retention. In particular, identifying customers who are about to churn (cancel their service) is potentially worth millions of dollars if the company can proactively address the reason that customer is considering cancellation and retain the customer. Data on past customers, some of whom churned and some who did not, have been collected. The variables in this data set are listed in the following table.

DATA*file* cellphonechurn.csv

Variable	Description
AccountWeeks	number of weeks customer has had active account
ContractRenewal	1 if customer recently renewed contract, 0 if not
DataPlan	1 if customer has data plan, 0 if not
DataUsage	gigabytes of monthly data usage
CustServCalls	number of calls into customer service
DayMins	average daytime minutes per month
DayCalls	average number of daytime calls
MonthlyCharge	average monthly bill
OverageFee	largest overage fee in last 12 months
RoamMins	average number of roaming minutes
Churn	"Yes" if customer cancelled service, "No" if not

Apply logistic regression with lasso regularization to classify observations as churning or not by using Churn as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 3, 8,9**

a. Determine the lasso regularization penalty that maximizes AUC in a validation procedure.

b. Apply the final model selected in part (a) to the test set. What are the values of the intercept and variable coefficients in this final model? Interpret the coefficients.

c. For the final model selected in part (a), compute and interpret the lift on the top 10% of test set observations most likely to churn.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

48. **Cellphone Customer Retention: *k*-Nearest Neighbors.** Refer to the scenario in Problem 47 regarding the identification of churning cellphone customers. Apply *k*-nearest neighbors to classify observations as churning or not by using Churn as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 4, 8**



    a. Based on all the input variables, determine the value of *k* that maximizes the AUC in a validation procedure.
    b. Remove DataPlan, RoamMins, OverageFee, and AccountWeeks from the set of input features considered and re-calibrate the value of *k* that maximizes the AUC. How does this *k*-nearest neighbors model compare to the model obtained in part (a)?
    c. For the best-performing *k*-nearest neighbors model in the validation procedure (with respect to AUC), compute and interpret the lift on the top 10% of test set observations most likely to churn.

49. **Cellphone Customer Retention: Classification Tree.** Refer to the scenario in Problem 47 regarding the identification of churning cellphone customers. Apply a classification tree to classify observations as churning or not by using Churn as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 5**



    a. Train a pruned classification tree and report its AUC from a validation experiment.
    b. Consider a customer described by: AccountWeeks = 40, ContractRenewal = 1, DataPlan = 1, DataUsage = 0, CustServCalls = 6, DayMins = 82, DayCalls = 123, MonthlyCharge = 32, OverageFee = 11, and RoamMins = 11. For a cutoff value of 0.5, does the pruned tree classify this observation as churning?
    c. For the best pruned tree identified in part (a), compute and interpret the lift on the top 10% of test set observations most likely to churn.

50. **Cellphone Customer Retention: Random Forest.** Refer to the scenario in Problem 47 regarding the identification of churning cellphone customers. Apply a random forest to classify observations as churning or not by using Churn as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 6**



    a. Determine the random forest configuration that maximizes the AUC in a validation procedure.
    b. For the best-performing random forest identified in part (a), compute and interpret the lift on the top 10% of test set observations most likely to churn.

51. **Cellphone Customer Retention: Neural Network.** Refer to the scenario in Problem 47 regarding the identification of churning cellphone customers. Apply a neural network to classify observations as churning or not by using Churn as the target (or response) variable. Set aside 20% of the data as a test set and use 80% of the data for training and validation. **LO 1, 2, 7**



    a. Determine the neural network configuration that maximizes the AUC in a validation procedure.
    b. For the best-performing neural network identified in part (a), compute and interpret the lift on the top 10% of test set observations most likely to churn.

52. **Best Picture Academy Award: Logistic Regression.** Each year, the American Academy of Motion Picture Arts and Sciences recognizes excellence in the film industry by honoring directors, actors, and writers with awards (called "Oscars") in different categories. The most notable of these awards is the Oscar for Best Picture. Data has been collected on a sample of movies nominated for the Best Picture Oscar. The variables include total number of Oscar nominations across all award categories, number of Golden Globe awards won (the Golden Globe award show precedes the Academy Awards), the genre of the movie, and whether or not the movie won the Best Picture Oscar award.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
oscars2022.csv

Variable	Description
Year	Year movie was nominated for Best Picture Oscar
Title	Movie title
OscarNominations	Total number of Oscar nominations across all award categories
GoldenGlobes	Total number of Golden Globe awards won
Genre	Movie genre
Winner	"Yes" if movie won Best Picture Oscar and "No" otherwise

Apply logistic regression with lasso regularization to classify winners of the Best Picture Oscar by using Winner as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 3, 8, 9**

a. Determine the lasso regularization penalty that maximizes AUC in a validation procedure.
b. For the level of lasso regularization identified in part (a), what are the values of the intercept and variable coefficients in this final model? Interpret the coefficients. Does the lasso regularization eliminate any variables?
c. Note that each year there is only one winner of the Best Picture Oscar (1 winner in each of the 38 years). Knowing this, examine the confusion matrix and explain what is wrong with classifying a movie based on a cutoff value?
d. What is the best way to use the model to predict the annual winner?

53. **Best Picture Academy Award: *k*-Nearest Neighbors.** Refer to the scenario in Problem 52 regarding the identification of movies that win the Best Picture Oscar. Apply *k*-nearest neighbors to classify observations as winning best picture or not by using Winner as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 4, 8**

DATA *file*
oscars2022.csv

a. Based on all the input variables, determine the value of *k* that maximizes the AUC in a validation procedure.
b. Note that each year there is only one winner of the Best Picture Oscar. Knowing this, examine the confusion matrix and explain what is wrong with classifying a movie based on a cutoff value?
c. What is the best way to use the model to predict the annual winner?

54. **Best Picture Academy Award: Classification Tree.** Refer to the scenario in Problem 52 regarding the identification of movies that win the Best Picture Oscar. Apply a classification tree to classify observations as winning best picture or not by using Winner as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 5**

DATA *file*
oscars2022.csv

a. Train a pruned classification tree and report its AUC from a validation experiment.
b. Consider a 2023 movie A described by: OscarNominations = 14, GoldenGlobes = 3, Genre = drama and 2023 movie B described by: OscarNominations = 8, GoldenGlobes = 7, and Genre = drama. For a cutoff value of 0.5, how does the pruned tree classify each of these movies? Knowing that there is only one winner of the Best Picture Oscar each year, what does this suggest about using a cutoff value to classify movies?
c. What is the best way to use the model to predict the annual winner?

55. **Best Picture Academy Award: Random Forest.** Refer to the scenario in Problem 52 regarding the identification of movies that win the Best Picture Oscar. Apply a random forest to classify observations as winning best picture or not by using Winner as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 6**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
oscars2022.csv

a. Determine the random forest configuration that maximizes the AUC in a validation procedure.
b. Note that each year there is only one winner of the Best Picture Oscar. Knowing this, examine the confusion matrix and explain what is wrong with classifying a movie based on a cutoff value?
c. What is the best way to use the model to predict the annual winner?

56. **Best Picture Academy Award: Neural Network.** Refer to the scenario in Problem 52 regarding the identification of movies that win the Best Picture Oscar. Apply a neural network to classify observations as winning best picture or not by using Winner as the target (or response) variable. Use 100% of the data for training and validation (do not use any data as a test set). **LO 1, 2, 7**

DATA*file*
oscars2022.csv

a. Determine the neural network configuration that maximizes the AUC in a validation procedure.
b. Note that each year there is only one winner of the Best Picture Oscar. Knowing this, examine the confusion matrix and explain what is wrong with classifying a movie based on a cutoff value?
c. What is the best way to use the model to predict the annual winner?

Case Problem: Grey Code Corporation

Grey Code Corporation (GCC) is a media and marketing company involved in magazine and book publishing and in television broadcasting. GCC's portfolio of home and family magazines has been a long-running strength, but it has expanded to become a provider of a spectrum of services such as market research, communications planning, web site advertising, etc., that can enhance its clients' brands.

GCC's relational database contains over a terabyte of data encompassing 75 million customers. GCC uses the data in its database to develop campaigns for new customer acquisition, customer reactivation, and identification of cross-selling opportunities for products. For example, GCC will generate separate versions of a monthly issue of a magazine that will differ only by the advertisements they contain. It will mail a subscribing customer the version with the print ads identified by its database as being of most interest to that customer.

DATA*file*
grey.csv
greylegend

One particular problem facing GCC is how to boost the customer response rate to renewal offers that it mails to its magazine subscribers. The industry response rate is about 2%, but GCC has historically performed better than that. However, GCC must update its model to correspond to recent changes. GCC's director of database marketing, Chris Grey, wants to make sure that GCC maintains its place as one of the top achievers in targeted marketing. The file *grey* contains 38 variables (columns) and over 40,000 rows (distinct customers). The table appended to the end of this case provides a list of the variables and their descriptions.

Play the role of Chris Grey and construct a classification model to identify customers who are likely to respond to a mailing. Write a report that documents the following steps: **LO 1, 2, 3, 4, 5, 6, 7, 8, 9**

1. Partition the data into a test set and data for training and validation.
2. Train and validate several candidate classification models.
3. Based on the performance in the validation experiment in part 2, select a final classification model and assess its performance on the test set.
4. Based on the final model's performance on the test set, report performance measures and relevant charts (ROC curves, lift charts, etc.) and include a recommendation on how to apply the results of your proposed model. For example, if GCC sends the targeted marketing to the top 10% of the test set that the model believes is most likely to renew, what is the expected response rate? How does that compare to the industry's average response rate?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Variable	Description
Renewal	1 if customer renewed magazine in response to mailing, 0 otherwise
Age	Customer age (ranges from 18 to 99)
HomeOwner	Likelihood of customer owning their own home
ResidenceLength	Number of years customer has lived at current residence. Values: 1 = less than two years, 2 = two years, 3 = three years, 4 = four years, 5 = five years, 6 = six years, 7 = seven years, 8 = eight years, 9 = nine years, 10 = ten years, 11 = eleven years, 12 = twelve years, 13 = thirteen years, 14 = fourteen years, or more
DwellingType	Identifies the type of residence. S = Single family dwelling unit. M = Multi family dwelling unit, U = unknown
Gender	F = female, M = male, U = unknown
Marital	S = Single, M = Married, O = Other (divorced, widowed, etc.), U = unknown
HouseholdSize	Identifies the number of individuals in the household. Arguments are: 1 = 1 person in the household, 2 = 2 people in the household, 3 = 3 people in the household, 4 = 4 people in the household, 5 = 5 people in the household, 6 = 6 or more people in the household
ChildPresent	Indicates if children are present in the home. Y = child 21 or younger present in the home; N = no child 21 or younger present in the home; U = unknown
Child0-5	Likelihood of child 0–5 years old present in home
Child6-12	Likelihood of child 6–12 years old present in home
Child13-18	Likelihood of child 13–18 years old present in home
Income	Estimated income. Ranges from $5,000 to $500,000+
Occupation	Broad aggregation of occupations into high level categories. Arguments are: R = retired, W = professional/executive, M = sales/marketing/services/clerical, B = skilled trades/laborers (blue collar type jobs), H = at home (caregivers, unemployed, homemakers), U = unknown
HomeValue	The estimated home value in ranges. Arguments are 1–10. 1 = under $50K, 2 = $50K—under $100K, 3 = $100K—under $150K, 4 = $150K—under $200K, 5 = $200K—under $250K, 6 = $250K—under $300K, 7 = $300K—under $350K, 8 = $350K—under $400K, 9 = over $400, 10 = Unknown
MagazineStatus	Identifies the status for a customer based on their magazine business activity. A = active subscriber, B = cancelled subscription due to non-payment, C = cancelled subscription, E = subscription expired within last 3 years, N = gift subscription, O = subscription expired over 3 years ago, S = subscription suspended at request of customer, U = unknown
PaidDirectMailOrders	Number of paid direct mail orders across all magazine subscriptions
YearsSinceLastOrder	Years since last order across all business lines
TotalAmountPaid	Total dollar amount paid for all magazine subscriptions over time

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Variable	Description
DollarsPerIssue	Paid Amount/Number of Issues Served. Average value per issue (takes the subscription term into account)
TotalPaidOrders	Total # of paid orders across all magazine subscriptions
MonthsSinceLastPayment	Recency - # months since most recent payment
LastPaymentType	Indicates how the customer paid on the most recent order. If it was credit order it will contain the billing effort number (how many bills were sent to collect payment). A = cash order, C = mass cancel for non-payment, D = Advanced renewal order cancelled, E = paid via collection agency, F = no billing, G = gift billed, I = online customer payment, K = payment received after order cancelled, L = extra payment, M = step-up financing, S = customer requested order cancellation, U = default paid, 0 = unpaid credit, 1 – 9 = paid credit on *i*th billing
UnpaidMagazines	Number of magazine titles currently in "unpaid" status for a given magazine customer
PaidCashMagazines	Number of magazine titles currently in "paid cash" status for a given magazine customer
PaidReinstateMagazines	Number of magazine titles currently in "paid reinstate" status for a given magazine customer
PaidCreditMagazines	Number of magazine titles currently in "paid credit" status for a given magazine customer
ActiveSubscriptions	Number of different magazines the customer is in "Active" status
ExpiredSubscriptions	Number of different magazines the customer is in "Expire" status
RequestedCancellations	Number of different magazines the customer is in "Cancelled via Customer Request" status
NoPayCancellations	Number of different magazines the customer is in "Cancelled for non-payment" status
PaidComplaints	Number of different magazines the customer is in "Paid Complaint" status
GiftDonor	Yes/No indicator as to whether the customer has given a magazine subscription as a gift
NumberGiftDonations	Number of subscription gift orders for this customer
MonthsSince1stOrder	Recency (in months) of first order for this magazine
MonthsSinceLastOrder	Recency (in months) of most recent order for this magazine
MonthsSinceExpire	Recency (in months) since the customer's subscription has expired for this magazine. Negative values represent months until an active subscription expires

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 12

Spreadsheet Models

Contents

Analytics in Action: Procter & Gamble

12.1 Building Good Spreadsheet Models
- Influence Diagrams
- Building a Mathematical Model
- Spreadsheet Design and Implementing the Model in a Spreadsheet

12.2 What-If Analysis
- Data Tables
- Goal Seek
- Scenario Manager

12.3 Some Useful Excel Functions for Modeling
- SUM and SUMPRODUCT
- IF and COUNTIF
- XLOOKUP

12.4 Auditing Spreadsheet Models
- Trace Precedents and Dependents
- Show Formulas
- Evaluate Formulas
- Error Checking
- Watch Window

12.5 Predictive and Prescriptive Spreadsheet Models

Summary 661
Glossary 661
Problems 662

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Construct an influence diagram of a problem and use it to construct a model.

LO 2 Build an Excel spreadsheet model of a business problem using formulas.

LO 3 Construct a one-way and two-way data table to quantify the impact of changing the value of one or two input variables on the value of an output variable.

LO 4 Conduct a breakeven analysis in Excel.

LO 5 Use Goal Seek to solve for the value of an input variable that will result in a specified value of an output variable.

LO 6 Use Scenario Manager to quantify the impact of changing the values of multiple input variables on the value of one or more output variables.

LO 7 Use Excel functions such as SUM, SUMPRODUCT, IF, COUNTIF, and XLOOKUP and other Excel functions in a spreadsheet model.

LO 8 Use Excel's Formula Auditing Tools to audit a spreadsheet model and correct errors.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

Procter & Gamble*

Procter & Gamble (P&G) is a Fortune 500 consumer goods company headquartered in Cincinnati, Ohio. P&G produces well-known brands such as Tide detergent, Gillette razors, Swiffer cleaning products, and many other consumer goods. P&G is a global company and has been recognized for its excellence in business analytics, including supply chain analytics and market research.

With operations around the world, P&G must do its best to maintain inventory at levels that meet its high customer service requirements. A lack of on-hand inventory can result in a stockout of a product and an inability to meet customer demand. This not only results in lost revenue for an immediate sale but can also cause customers to switch permanently to a competing brand. On the other hand, excessive inventory forces P&G to invest cash in inventory when that money could be invested in other opportunities, such as research and development.

To ensure that the inventory of its products around the world is set at appropriate levels, P&G analytics personnel developed and deployed a series of spreadsheet inventory models. These spreadsheets implement mathematical inventory models to tell business units when and how much to order to keep inventory levels where they need to be to maintain service and keep investment as low as possible.

The spreadsheet models were carefully designed to be easily understood by the users and easy to use and interpret. Their users can also customize the spreadsheets to their individual situations.

Over 70% of the P&G business units use these models, with a conservative estimate of a 10% reduction in inventory around the world. This equates to a cash savings of nearly $350 million.

*I. Farasyn, K. Perkoz, and W. Van de Velde, "Spreadsheet Model for Inventory Target Setting at Procter & Gamble," *Interfaces* 38, no. 4 (July–August 2008): 241–250.

Numerous specialized software packages are available for descriptive, predictive, and prescriptive business analytics. Because these software packages are specialized, they usually provide the user with numerous options and the capability to perform detailed analyses. However, they tend to be considerably more expensive than a spreadsheet package such as Excel. Also, specialized packages often require substantial user training. Because spreadsheets are less expensive, often come preloaded on computers, and are fairly easy to use, they are without question the most-used business analytics tool. Every day, millions of people around the world use spreadsheet decision models to perform risk analysis, inventory tracking and control, investment planning, breakeven analysis, and many other essential business planning and decision tasks. A well-designed, well-documented, and accurate spreadsheet model can be a very valuable tool in decision making.

If you are unfamiliar with spreadsheets, we suggest you first review the material in Appendix A. It provides basic information that is fundamental to using Excel.

Spreadsheet models are mathematical and logic-based models. Their strength is that they provide easy-to-use, sophisticated mathematical and logical functions, allowing for easy instantaneous recalculation for a change in model inputs. This is why spreadsheet models are often referred to as **what-if models**. What-if models allow you to answer questions such as, "If the per unit cost is $4, what is the impact on profit?" Changing data in a given cell has an impact not only on that cell but also on any other cells containing a formula or function that uses that cell.

In this chapter, we discuss principles for building reliable spreadsheet models. We begin with a discussion of how to build a conceptual model of a decision problem, how to convert the conceptual model to a mathematical model, and how to implement the model in a spreadsheet. We introduce three analysis tools available in Excel: Data Tables, Goal Seek, and Scenario Manager. We discuss some Excel functions that are useful for building spreadsheet models for decision making. Finally, we present how to audit a spreadsheet model to ensure its reliability.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

12.1 Building Good Spreadsheet Models

Let us begin our discussion of spreadsheet models by considering the cost of producing a single product. The total cost of manufacturing a product can usually be defined as the sum of two costs: fixed cost and variable cost. *Fixed cost* is the portion of the total cost that does not depend on the production quantity; this cost remains the same no matter how much is produced. *Variable cost*, on the other hand, is the portion of the total cost that is dependent on and varies with the production quantity. To illustrate how cost models can be developed, we will consider a manufacturing problem faced by Nowlin Plastics.

Nowlin Plastics produces a line of cell phone covers. Nowlin's best-selling cover is its Viper model, a slim but very durable black and gray plastic cover. The annual fixed cost to produce the Viper cover is $234,000. This fixed cost includes management time and other costs that are incurred regardless of the number of units eventually produced. In addition, the total variable cost, including labor and material costs, is $2 for each unit produced.

Nowlin is considering outsourcing the production of some products for next year, including the Viper. Nowlin has a bid from an outside firm to produce the Viper for $3.50 per unit. Although it is more expensive per unit to outsource the Viper ($3.50 versus $2.00), the fixed cost can be avoided if Nowlin purchases rather than manufactures the product. Next year's exact demand for Viper is not yet known. Nowlin would like to compare the costs of manufacturing the Viper in-house to those of outsourcing its production to another firm, and management would like to do that for various production quantities. Many manufacturers face this type of decision, which is known as a **make-versus-buy decision**.

Influence Diagrams

It is often useful to begin the modeling process with a conceptual model that shows the relationships between the various parts of the problem being modeled. The conceptual model helps in organizing the data requirements and provides a road map for eventually constructing a mathematical model. A conceptual model also provides a clear way to communicate the model to others. An **influence diagram** is a visual representation of which entities influence others in a model. Parts of the model are represented by circular or oval symbols called *nodes*, and arrows connecting the nodes show influence.

Figure 12.1 shows an influence diagram for Nowlin's total cost of production for the Viper. Total manufacturing cost depends on fixed cost and variable cost, which in turn depends on the variable cost per unit and the quantity required.

An expanded influence diagram that includes an outsourcing option is shown in Figure 12.2. Note that the influence diagram in Figure 12.1 is a subset of the influence diagram in Figure 12.2. Our method here—namely, to build an influence diagram for a portion of the problem and then expand it until the total problem is conceptually modeled—is usually a good way to proceed. This modular approach simplifies the process and reduces the likelihood of error. This is true not just for influence diagrams but for the construction of the mathematical and spreadsheet models as well. Next we turn our attention to using the influence diagram in Figure 12.2 to guide us in the construction of the mathematical model.

Building a Mathematical Model

The task now is to use the influence diagram to build a mathematical model. Let us first consider the cost of manufacturing the required units of the Viper. As the influence diagram shows, this cost is a function of the fixed cost, the variable cost per unit, and the quantity required. In general, it is best to define notation for every node in the influence diagram. Let us define the following:

$$\begin{aligned} q &= \text{quantity (number of units) required} \\ FC &= \text{the fixed cost of manufacturing} \\ VC &= \text{the per-unit variable cost of manufacturing} \\ TMC(q) &= \text{total cost to manufacture } q \text{ units} \end{aligned}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.1 An Influence Diagram for Nowlin's Manufacturing Cost

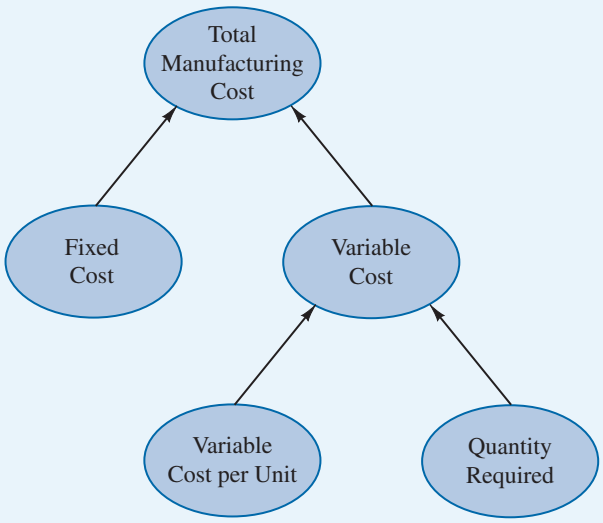
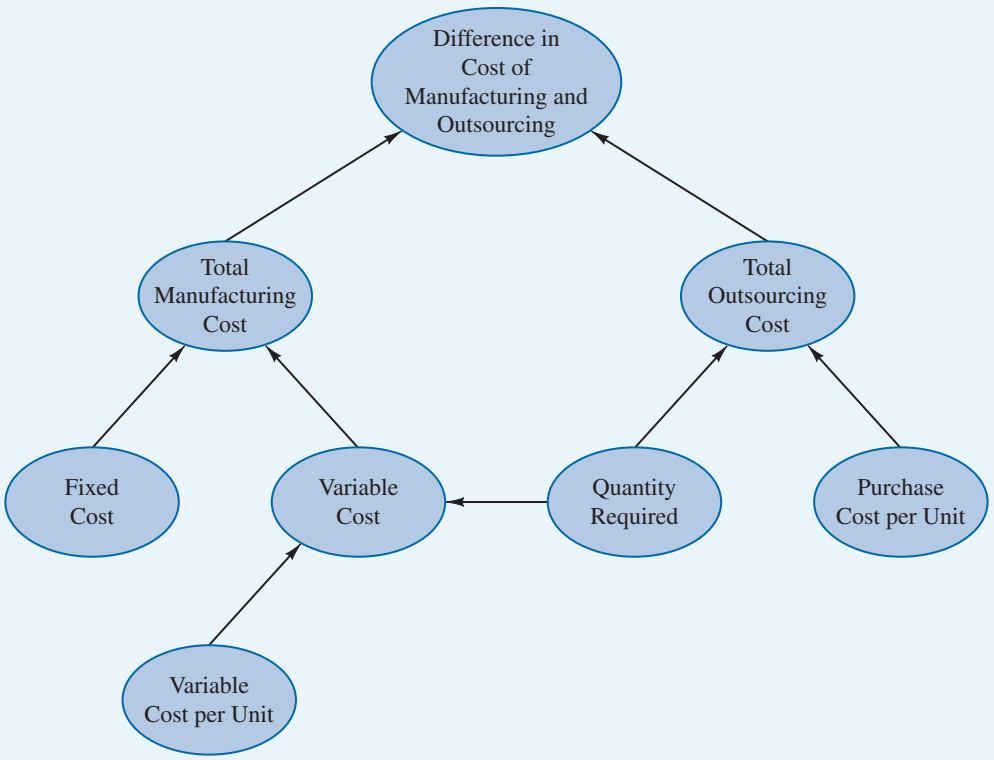


Figure 12.2 An Influence Diagram for Comparing Manufacturing versus Outsourcing Cost for Nowlin Plastics



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The cost-volume model for producing q units of the Viper can then be written as follows:

$$TMC(q) = FC + (VC \times q) \quad \textbf{(12.1)}$$

For the Viper, $FC = \$234{,}000$ and $VC = \$2$, so that equation (12.1) becomes

$$TMC(q) = \$234{,}000 + \$2q$$

Once a quantity required (q) is established, equation (12.1), now populated with the data for the Viper, can be used to compute the total manufacturing cost. For example, the decision to produce $q = 10{,}000$ units would result in a total manufacturing cost of $TMC(10{,}000) = \$234{,}000 + \$2(10{,}000) = \$254{,}000$.

Similarly, a mathematical model for purchasing q units is shown in equation (12.2). Let $P =$ the per-unit purchase cost and $TPC(q) =$ the total cost to outsource or purchase q units:

$$TPC(q) = Pq \quad \textbf{(12.2)}$$

For the Viper, since $P = \$3.50$, equation (12.2) becomes

$$TPC(q) = \$3.5q$$

Thus, the total cost to outsource 10,000 units of the Viper is $TPC(10{,}000) = 3.5(10{,}000) = \$35{,}000$.

We can now state mathematically the savings associated with outsourcing. Let $S(q) =$ the savings due to outsourcing, that is, the difference between the total cost of manufacturing q units and the total cost of purchasing q units:

$$S(q) = TMC(q) - TPC(q) \quad \textbf{(12.3)}$$

In summary, Nowlin's decision problem is whether to manufacture or outsource the demand for its Viper product next year. Because management does not yet know the required demand, the key question is, "For what quantities is it more cost-effective to outsource rather than manufacture the Viper?" Mathematically, this question is, "For what values of q is $S(q) > 0$?" Next we discuss a spreadsheet implementation of our conceptual and mathematical models that will help us answer this question.

Spreadsheet Design and Implementing the Model in a Spreadsheet

There are several guiding principles for how to build a spreadsheet so that it is easily used by others and the risk of error is mitigated. In this section, we discuss some of those principles and illustrate the design and construction of a spreadsheet model using the Nowlin Plastics make-versus-buy decision.

In the construction of a spreadsheet model, it is helpful to categorize its components. For the Nowlin Plastics problem, we have defined the following components (corresponding to the nodes of the influence diagram in Figure 12.2):

q = number of units required
FC = the fixed cost of manufacturing
VC = the per-unit variable cost of manufacturing
$TMC(q)$ = total cost to manufacture q units
P = the per-unit purchase cost
$TPC(q)$ = the total cost to purchase q units
$S(q)$ = the savings from outsourcing q units

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note that q, FC, VC, *and* P *each is the beginning of a path in the influence diagram in Figure 12.2. In other words, they have no inward-pointing arrows.*

Several points are in order. Some of these components are a function of other components (*TMC*, *TPC*, and *S*), and some are not (*q*, *FC*, *VC*, and *P*). *TMC*, *TPC*, and *S* will be formulas involving other cells in the spreadsheet model, whereas *q*, *FC*, *VC*, and *P* will just be entries in the spreadsheet. Furthermore, the value we can control or choose is *q*. In our analysis, we seek the value of *q*, such that $S(q) > 0$; that is, the savings associated with outsourcing is positive. The number of Vipers to make or buy for next year is Nowlin's decision. So we will treat *q* somewhat differently than *FC*, *VC,* and *P* in the spreadsheet model, and we refer to the quantity *q* as a **decision variable**. *FC*, *VC*, and *P* are measurable factors that define characteristics of the process we are modeling and so are *uncontrollable inputs* to the model, which we refer to as **parameters** of the model.

Figure 12.3 shows a spreadsheet model for the Nowlin Plastics make-versus-buy decision.

Column A is reserved for labels, including cell A1, where we have named the model "Nowlin Plastics." The input parameters (*FC*, *VC*, and *P*) are placed in cells B4, B5, and B7, respectively. We offset *P* from *FC* and *VC* because it is for outsourcing. We have created a parameters section in the upper part of the sheet. Below the parameters section, we have created the Model section. The first entry in the Model section is the quantity *q*—the number of units of Viper produced or purchased in cell B11—and shaded it to signify that this is a decision variable. We have placed the formulas corresponding to equations (12.1) to (12.3)

Figure 12.3 Nowlin Plastics Make-versus-Buy Spreadsheet Model

	A	B	C
1	**Nowlin Plastics**		
2			
3	**Parameters**		
4	Manufacturing Fixed Cost	234000	
5	Manufacturing Variable Cost per Unit	2	
6			
7	Outsourcing Cost per Unit	3.5	
8			
9			
10	**Model**		
11	Quantity	10000	
12			
13	Total Cost to Produce	=B4+B11*B5	
14			
15	Total Cost to Outsource	=B7*B11	
16			
17	Savings due to Outsourcing	=B13-B15	
18			

MODEL *file*
nowlin

	A	B	C
1	**Nowlin Plastics**		
2			
3	**Parameters**		
4	Manufacturing Fixed Cost	$234,000.00	
5	Manufacturing Variable Cost per Unit	$2.00	
6			
7	Outsourcing Cost per Unit	$3.50	
8			
9			
10	**Model**		
11	Quantity	10,000	
12			
13	Total Cost to Produce	$254,000.00	
14			
15	Total Cost to Outsource	$35,000.00	
16			
17	Savings due to Outsourcing	$219,000.00	
18			

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

As described in Appendix A, Excel formulas always begin with an equal sign.

in cells B13, B15, and B17. Cell B13 corresponds to equation (12.1), cell B15 to (12.2), and cell B17 to (12.3).

In cell B11 of Figure 12.3, we have set the value of q to 10,000 units. The model shows that the cost to manufacture 10,000 units is \$254,000, the cost to purchase the 10,000 units is \$35,000, and the savings from outsourcing is \$219,000. At a quantity of 10,000 units, we see that it is better to incur the higher variable cost (\$3.50 versus \$2) than to manufacture and have to incur the additional fixed cost of \$234,000. It will take a value of q larger than 10,000 units to make up the fixed cost incurred when Nowlin manufactures the product. At this point, we could increase the value of q by placing a value higher than 10,000 in cell B11 and see how much the savings in cell B17 decreases, doing this until the savings are close to zero. This is called a *trial-and-error approach*. Fortunately, Excel has what-if analysis tools that will help us use our model to further analyze the problem. We will discuss these what-if analysis tools in Section 12.2. Before doing so, let us first review what we have learned in constructing the Nowlin spreadsheet model.

The general principles of spreadsheet model design and construction are as follows:

- Separate the parameters from the model.
- Document the model, and use proper formatting and color as needed.
- Use simple formulas.

Let us discuss the general merits of each of these points.

Separate the Parameters from the Model Separating the parameters from the model enables the user to update the model parameters without the risk of mistakenly creating an error in a formula. For this reason, it is good practice to have a parameters section at the top of the spreadsheet. A separate model section should contain all calculations. For a what-if model or an optimization model, some cells in the model section might also correspond to controllable inputs or decision variables (values that are not parameters or calculations but are the values we choose). The Nowlin model in Figure 12.3 is an example of this. The parameters section is in the upper part of the spreadsheet, followed by the model section, below which are the calculations and a decision cell (B11 for q in our model). Cell B11 is shaded to signify that it is a decision cell.

Document the Model and Use Proper Formatting and Color as Needed A good spreadsheet model is well documented. Clear labels and proper formatting and alignment facilitate navigation and understanding. For example, if the values in a worksheet are cost, currency formatting should be used. Also, no cell with content should be unlabeled. A new user should be able to easily understand the model and its calculations. If color makes a model easier to understand and navigate, use it for cells and labels.

Use Simple Formulas Clear, simple formulas can reduce errors and make it easier to maintain the spreadsheet. Long and complex calculations should be divided into several cells. This makes the formula easier to understand and easier to edit. Avoid using numbers in a formula (separate the data from the model). Instead, put the number in a cell in the parameters section of your worksheet and refer to the cell location in the formula. Building the formula in this manner avoids having to edit the formula for a simple data change. For example, equation (12.3), the savings due to outsourcing, can be calculated as follows: $S(q) = TMC(q) - TPC(q) = FC + (VC)q - Pq = FC + (VC - P)q$. Since $VC - P = 3.50 - 2 = 1.50$, we could have just entered the following formula in a single cell: $=234{,}000 - 1.50 * \text{B11}$. This is a very bad idea because if any of the input data change, the formula must be edited. Furthermore, the user would not know the values of VC and P, only that, for the current values, the difference is 1.50. The approach in Figure 12.3 is more transparent, is simpler, lends itself better to analysis of changes in the parameters, and is less likely to contain errors.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notes + Comments

1. Some users of influence diagrams recommend using different symbols for the various types of model entities. For example, circles might denote known inputs, ovals might denote uncertain inputs, rectangles might denote decisions or controllable inputs, triangles might denote calculations, and so forth.
2. The use of color in a spreadsheet model is an effective way to draw attention to a cell or set of cells. For example, we shaded cell B11 in Figure 12.3 to draw attention to the fact that q is a controllable input. However, avoid using too much color which can overwhelm users and actually have a negative impact on their ability to understand the model.
3. Clicking the **Formulas** tab in the Ribbon of Excel and then selecting **Show Formulas** from the **Formula Auditing** area will display all formulas used in an Excel spreadsheet.

12.2 What-If Analysis

Excel offers a number of tools to facilitate what-if analysis. In this section, we introduce three such tools: Data Tables, Goal Seek, and Scenario Manager. All of these tools are designed to rid the user of the tedious manual trial-and-error approach to analysis. Let us see how each of these tools can be used to aid with what-if analysis.

Data Tables

An Excel **data table** quantifies the impact of changing the value of a specific input on an output of interest. Excel can generate either a **one-way data table**, which summarizes a single input's impact on the output, or a **two-way data table**, which summarizes two inputs' impact on the output.

Let us continue with our Nowlin Plastics example and consider how savings due to outsourcing changes as the quantity of Vipers changes. This should help us answer the question, "For which values of q is outsourcing more cost-effective?" A one-way data table changing the value of quantity and reporting savings due to outsourcing would be very useful. We will use the previously developed Nowlin spreadsheet for this analysis.

The first step in creating a one-way data table is to construct a sorted list of the values you would like to consider for the input. Let us investigate the quantity q over a range from 0 to 300,000 in increments of 25,000 units. Figure 12.4 shows the data entered in cells D5 through D17, with a column label in D4. This column of data is the set of values that Excel will use as inputs for q. Since the output of interest is savings due to outsourcing (located in cell B17), we have entered the formula =B17 in cell E4. In general, set the cell to the right of the label to the cell location of the output variable of interest. Once the basic structure is in place, we invoke the Data Table tool using the following steps:

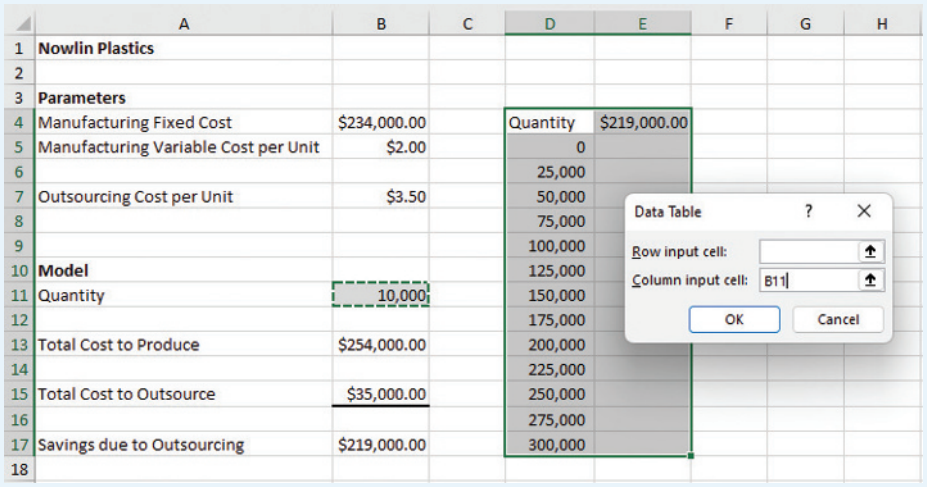
Entering B11 *in the* ***Column input cell:*** *box indicates that the column of data corresponds to different values of the input located in cell B11.*

Step 1. Select cells D4:E17
Step 2. Click the **Data** tab in the Ribbon
Step 3. Click **What-If Analysis** in the **Forecast** group, and select **Data Table**
Step 4. When the **Data Table** dialog box appears:
Enter *B11* in the **Column input cell:** box
Click **OK**

As shown in Figure 12.5, the table will be populated with the value of savings due to outsourcing for each value of quantity of Vipers in the table. For example, when $q = 25{,}000$ we see that $S(25{,}000) = \$196{,}500$, and when $q = 250{,}000$, $S(250{,}000) = -\$141{,}000$. A negative value for savings due to outsourcing means that manufacturing is cheaper than outsourcing for that quantity.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.4 The Input for Constructing a One-Way Data Table for Nowlin Plastics



	A	B	C	D	E	F	G	H
1	**Nowlin Plastics**							
2								
3	**Parameters**							
4	Manufacturing Fixed Cost	$234,000.00		Quantity	$219,000.00			
5	Manufacturing Variable Cost per Unit	$2.00		0				
6				25,000				
7	Outsourcing Cost per Unit	$3.50		50,000				
8				75,000				
9				100,000				
10	**Model**			125,000				
11	Quantity	10,000		150,000				
12				175,000				
13	Total Cost to Produce	$254,000.00		200,000				
14				225,000				
15	Total Cost to Outsource	$35,000.00		250,000				
16				275,000				
17	Savings due to Outsourcing	$219,000.00		300,000				
18								

Figure 12.5 Results of One-Way Data Table for Nowlin Plastics

	A	B	C	D	E
1	**Nowlin Plastics**				
2					
3	**Parameters**				
4	Manufacturing Fixed Cost	$234,000.00		Quantity	$219,000.00
5	Manufacturing Variable Cost per Unit	$2.00		0	$234,000
6				25,000	$196,500
7	Outsourcing Cost per Unit	$3.50		50,000	$159,000
8				75,000	$121,500
9				100,000	$84,000
10	**Model**			125,000	$46,500
11	Quantity	10,000		150,000	$9,000
12				175,000	-$28,500
13	Total Cost to Produce	$254,000.00		200,000	-$66,000
14				225,000	-$103,500
15	Total Cost to Outsource	$35,000.00		250,000	-$141,000
16				275,000	-$178,500
17	Savings due to Outsourcing	$219,000.00		300,000	-$216,000
18					

We have learned something very valuable from this table. Not only have we quantified the savings due to outsourcing for a number of quantities, we know too that for quantities of 150,000 units or less, outsourcing is cheaper than manufacturing and for quantities of 175,000 units or more, manufacturing is cheaper than outsourcing. Depending on Nowlin's confidence in their demand forecast for the Viper product for next year, we have likely

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

satisfactorily answered the make-versus-buy question. If, for example, management is highly confident that demand will be at least 200,000 units of Viper, then clearly they should manufacture the Viper rather than outsource. If management believes that Viper demand next year will be close to 150,000 units, they might still decide to manufacture rather than outsource. At 150,000 units, the savings due to outsourcing is only $9,000. That might not justify outsourcing if, for example, the quality assurance standards at the outsource firm are not at an acceptable level. We have provided management with valuable information that they may use to decide whether to make or buy. Next we illustrate how to construct a two-way data table.

Suppose that Nowlin has now received five different bids on the per-unit cost for outsourcing the production of the Viper. Clearly, the lowest bid provides the greatest savings. However, the selection of the outsource firm—if Nowlin decides to outsource—will depend on many factors, including reliability, quality, and on-time delivery. So it would be instructive to quantify the differences in savings for various quantities and bids. The five current bids are $2.89, $3.13, $3.50, $3.54, and $3.59. We may use the Excel Data Table to construct a two-way data table with quantity as a column and the five bids as a row, as shown in Figure 12.6.

In Figure 12.6, we have entered various quantities in cells D5 through D17, as in the one-way table. These correspond to cell B11 in our model. In cells E4 through I4, we have entered the bids. These correspond to B7, the outsourcing cost per unit. In cell D4, above the column input values and to the left of the row input values, we have entered the formula =B17, the location of the output of interest, in this case, savings due to outsourcing. Once the table inputs have been entered into the spreadsheet, we perform the following steps to construct the two-way data table.

Step 1. Select cells D4:I17
Step 2. Click the **Data** tab in the Ribbon
Step 3. Click **What-If Analysis**  in the **Forecast** group, and select **Data Table**
Step 4. When the **Data Table** dialog box appears:
Enter *B7* in the **Row input cell:** box
Enter *B11* in the **Column input cell:** box
Click **OK**

Figure 12.6 shows the selected cells and the Data Table dialog box. The results are shown in Figure 12.7.

We now have a table that shows the savings due to outsourcing for each combination of quantity and bid price. For example, for 75,000 Vipers at a cost of $3.13 per unit, the savings from buying versus manufacturing the units is $149,250. We can also see the range for the quantity for each bid price that results in a negative savings. For these quantities and bid combinations, it is better to manufacture than to outsource.

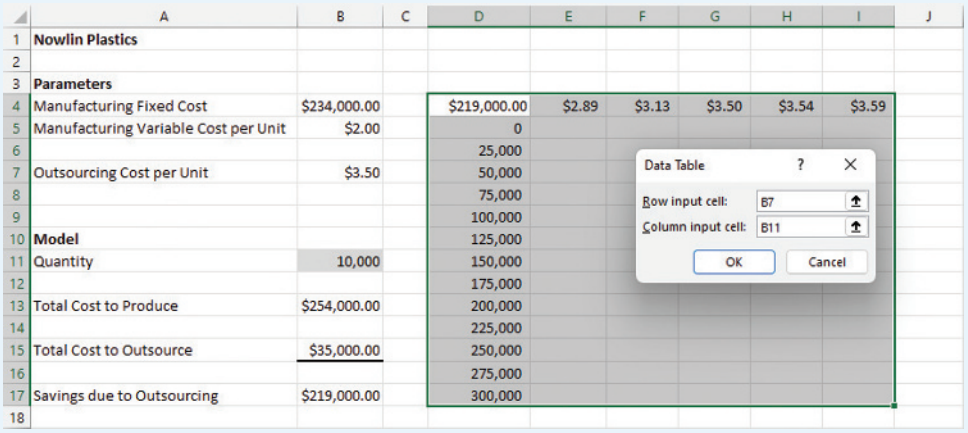
Using the Data Table allows us to quantify the savings due to outsourcing for the quantities and bid prices specified. However, the table does not tell us the exact number at which the transition occurs from outsourcing being cheaper to manufacturing being cheaper. For example, although it is clear from the table that for a bid price of $3.50, the savings due to outsourcing goes from positive to negative at some quantity between 150,000 units and 175,000 units, we know only that this transition occurs somewhere in that range. As we illustrate next, the what-if analysis tool Goal Seek can tell us the precise number at which this transition occurs.

Goal Seek

Excel's **Goal Seek** tool allows the user to determine the value of an input cell that will cause the value of a related output cell to equal some specified value (the *goal*). In the case

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.6 The Input for Constructing a Two-Way Data Table for Nowlin Plastics



	A	B	C	D	E	F	G	H	I	J
1	**Nowlin Plastics**									
2										
3	**Parameters**									
4	Manufacturing Fixed Cost	$234,000.00		$219,000.00	$2.89	$3.13	$3.50	$3.54	$3.59	
5	Manufacturing Variable Cost per Unit	$2.00		0						
6				25,000						
7	Outsourcing Cost per Unit	$3.50		50,000						
8				75,000						
9				100,000						
10	**Model**			125,000						
11	Quantity	10,000		150,000						
12				175,000						
13	Total Cost to Produce	$254,000.00		200,000						
14				225,000						
15	Total Cost to Outsource	$35,000.00		250,000						
16				275,000						
17	Savings due to Outsourcing	$219,000.00		300,000						
18										

Figure 12.7 Results of a Two-Way Data Table for Nowlin Plastics

	A	B	C	D	E	F	G	H	I
1	**Nowlin Plastics**								
2									
3	**Parameters**								
4	Manufacturing Fixed Cost	$234,000.00		$219,000.00	$2.89	$3.13	$3.50	$3.54	$3.59
5	Manufacturing Variable Cost per U	$2.00		0	$234,000	$234,000	$234,000	$234,000	$234,000
6				25,000	$211,750	$205,750	$196,500	$195,500	$194,250
7	Outsourcing Cost per Unit	$3.50		50,000	$189,500	$177,500	$159,000	$157,000	$154,500
8				75,000	$167,250	$149,250	$121,500	$118,500	$114,750
9				100,000	$145,000	$121,000	$84,000	$80,000	$75,000
10	**Model**			125,000	$122,750	$92,750	$46,500	$41,500	$35,250
11	Quantity	10,000		150,000	$100,500	$64,500	$9,000	$3,000	-$4,500
12				175,000	$78,250	$36,250	-$28,500	-$35,500	-$44,250
13	Total Cost to Produce	$254,000.00		200,000	$56,000	$8,000	-$66,000	-$74,000	-$84,000
14				225,000	$33,750	-$20,250	-$103,500	-$112,500	-$123,750
15	Total Cost to Outsource	$35,000.00		250,000	$11,500	-$48,500	-$141,000	-$151,000	-$163,500
16				275,000	-$10,750	-$76,750	-$178,500	-$189,500	-$203,250
17	Savings due to Outsourcing	$219,000.00		300,000	-$33,000	-$105,000	-$216,000	-$228,000	-$243,000
18									

of Nowlin Plastics, suppose we want to know the value of the quantity of Vipers at which it becomes more cost-effective to manufacture rather than outsource. For example, we see from the table in Figure 12.7 that for a bid price of $3.50 and some quantity between 150,000 units and 175,000 units, savings due to outsourcing goes from positive to negative. Somewhere in this range of quantity, the savings due to outsourcing is zero, and that is the point at which Nowlin would be indifferent to manufacturing and outsourcing. We may use Goal Seek to find the quantity of Vipers that satisfies the goal of zero savings due to

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

outsourcing for a bid price of $3.50. The following steps describe how to use Goal Seek to find this point.

Step 1. Click the **Data** tab in the Ribbon
Step 2. Click **What-If Analysis** in the **Forecast** group, and select **Goal Seek**
Step 3. When the **Goal Seek** dialog box appears (Figure 12.8):
Enter *B17* in the **Set cell:** box
Enter *0* in the **To value**: box
Enter *B11* in the **By changing cell:** box
Click **OK**
Step 4. When the **Goal Seek Status** dialog box appears, click **OK**

The completed Goal Seek dialog box is shown in Figure 12.8.

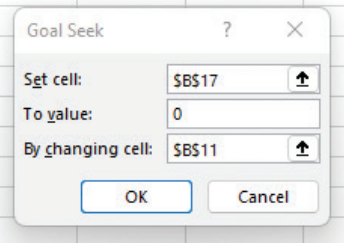
The results from Goal Seek are shown in Figure 12.9. The savings due to outsourcing in cell B17 is $0.00, and the quantity in cell B11 has been set by Goal Seek to 156,000. When the annual quantity required is 156,000, it costs $564,000 either to manufacture the product or to purchase it. We have already seen that lower values of the quantity required favor outsourcing. Beyond the value of 156,000 units it becomes cheaper to manufacture the product.

Scenario Manager

As we have seen, data tables are useful for exploring the impact of changing one or two model inputs on a model output of interest. **Scenario Manager** is an Excel tool that quantifies the impact of changing multiple inputs (a setting of these multiple inputs is called a scenario) on one or more outputs of interest. That is, Scenario Manager extends the data table concept to cases when you are interested in changing more than two inputs and want to quantify the changes these inputs have on one or more outputs of interest.

Figure 12.8 Goal Seek Dialog Box for Nowlin Plastics

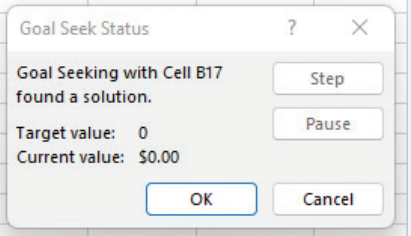
	A	B	C	D	E	F
1	**Nowlin Plastics**					
2						
3	**Parameters**					
4	Manufacturing Fixed Cost	$234,000.00				
5	Manufacturing Variable Cost per Unit	$2.00				
6						
7	Outsourcing Cost per Unit	$3.50				
8						
9						
10	**Model**					
11	Quantity	10,000				
12						
13	Total Cost to Produce	$254,000.00				
14						
15	Total Cost to Outsource	$35,000.00				
16						
17	Savings due to Outsourcing	$219,000.00				
18						



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.9 Results from Goal Seek for Nowlin Plastics

	A	B	C	D	E	F
1	**Nowlin Plastics**					
2						
3	**Parameters**					
4	Manufacturing Fixed Cost	$234,000.00				
5	Manufacturing Variable Cost per Unit	$2.00				
6						
7	Outsourcing Cost per Unit	$3.50				
8						
9						
10	**Model**					
11	Quantity	156,000				
12						
13	Total Cost to Produce	$546,000.00				
14						
15	Total Cost to Outsource	$546,000.00				
16						
17	Savings due to Outsourcing	$0.00				
18						



To illustrate the use of Scenario Manager, let us consider the case of the Middletown Amusement Park. John Miller, the manager at Middletown, has developed a simple spreadsheet model of the park's daily profit. Miller's model is shown in Figure 12.10.

On any given day, there are two types of customers in the park, those who own season passes and those who do not. Season-pass owners pay an annual membership fee during the offseason, but then pay nothing at the gate to enter the park. Those who are not season pass holders pay $35 per person to enter the park for the day. John refers to these non-season-pass holders as "admissions." On average, a season-pass holder spends $15 per person in the park on food, drinks, and novelties and an admission spends on average $45. The average daily cost of operations (including fixed costs) is $33,000 per day and the cost of goods is 50% of the price of the good. These data are reflected in John's spreadsheet model, which calculates a daily profit. As shown in Figure 12.10, for the data just described, John's model calculates the profit to be $81,500.

As you might expect, the profit generated on any given day is very dependent on the weather. As shown in Table 12.1, John has developed three weather-based scenarios: Partly Cloudy, Rain, and Sunny. The weather has a direct impact on four input parameters: the number of season-pass holders who enter the park, the number of non-season-pass holders (admissions) who enter the park, the amount each of these groups spends on average and the cost of operations. The Scenario Manager allows us to generate a report that gives an output variable or set of output variables of interest for each scenario. In this case, Scenario Manager will provide a report that gives the profit for each scenario.

The following steps describe how to use Scenario Manager to generate a scenario summary report.

Step 1. Click the **Data** tab in the Ribbon.
Step 2. Click **What-if Analysis** in the **Forecast** group and select **Scenario Manager**...

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.10 Middletown Amusement Park Daily Profit Model

	A	B	C
1	**Middletown Amusement Park**		
2			
3	**Parameters**		
4			
5	Admission Price	35	
6	Number of Season-Pass Holders Admitted	3000	
7	Admissions	1600	
8	Average Expenditure - Season Pass Holders	15	
9	Average Expenditure - Admissions	45	
10			
11	Cost of Operations	33000	
12	Cost of Goods %	0.5	
13			
14	**Model**		
15			
16	Admissions Revenue	=B5*B7	
17	Season Pass Holder Expenditures Revenue	=B6*B8	
18	Admissions Expenditures Revenue	=B7*B9	
19	Total Revenue	=B16+B17+B18	
20			
21	Cost of Operations	=B11	
22	Cost of Goods	=B12*(B17+B18)	
23	Total Cost	=B21+B22	
24			
25	Profit	=B19-B23	
26			

MODEL file
middletown

	A	B	C
1	**Middletown Amusement Park**		
2			
3	**Parameters**		
4			
5	Admission Price	$35	
6	Number of Season-Pass Holders Admitted	3000	
7	Admissions	1600	
8	Average Expenditure - Season Pass Holders	$15	
9	Average Expenditure - Admissions	$45	
10			
11	Cost of Operations	$33,000	
12	Cost of Goods %	50%	
13			
14	**Model**		
15			
16	Admissions Revenue	$56,000	
17	Season Pass Holder Expenditures Revenue	$45,000	
18	Admissions Expenditures Revenue	$72,000	
19	Total Revenue	$173,000	
20			
21	Cost of Operations	$33,000	
22	Cost of Goods	$58,500	
23	Total Cost	$91,500	
24			
25	Profit	$81,500	
26			

Step 3. When the **Scenario Manager** dialog box appears (Figure 12.11), click the **Add**... button

Step 4. When the **Add Scenario** dialog box appears (Figure 12.12):
Enter *Partly Cloudy* in the **Scenario name**: box
Enter *$B$6:$B$9,$B$11* in the **Changing cells**: box
Click **OK**

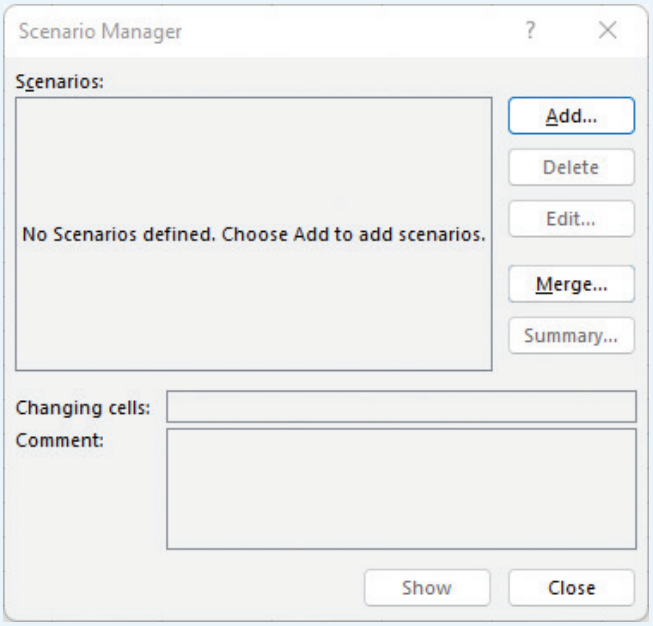
Step 5. When the **Scenario Values** dialog box appears (Figure 12.13):
Enter *3000* in the **$B$6** box
Enter *1600* in the **$B$7** box
Enter *15* in the **$B$8** box
Enter *45* in the **$B$9** box
Enter *33000* in the **$B$11** box
Click **OK**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 12.1 Weather Scenarios for Middletown Amusement Park

	Scenarios		
	Partly Cloudy	**Rain**	**Sunny**
Season-pass Holders	3,000	1,200	8,000
Admissions	1,600	250	2,400
Average Expenditure - Season-Pass Holders	$15	$10	$18
Average Expenditure - Admissions	$45	$20	$57
Cost of Operations	$33,000	$27,000	$37,000

Figure 12.11 Scenario Manager Dialog Box



Step 6. When the **Scenario Manager** dialog box appears, repeat steps 3–5 for each scenario shown in Table 12.1 (Rain and Sunny)

Step 7. When all scenarios have been entered and the **Scenario Manager** dialog box appears, click **Summary...**

Step 8. When the **Scenario Summary** dialog box appears (Figure 12.14):
Select **Scenario summary**
Enter *B25* in the **Result Cells** box
Click **OK**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.12 Add Scenario Dialog Box

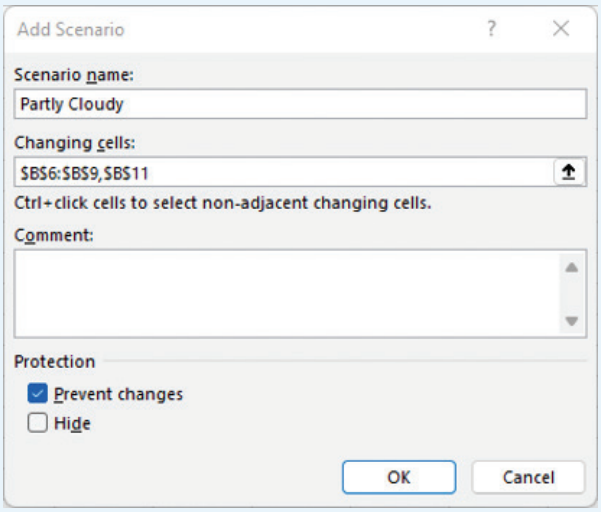


Figure 12.13 Scenario Values Dialog Box

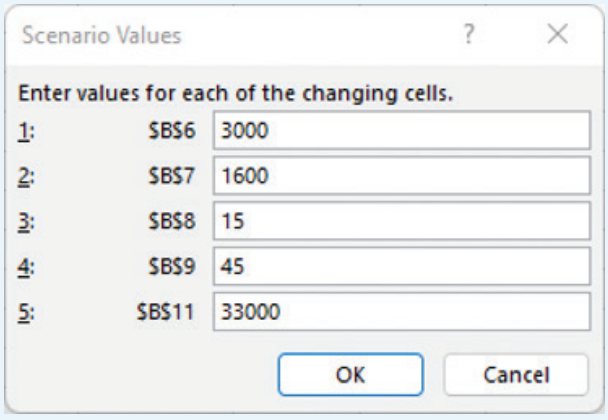
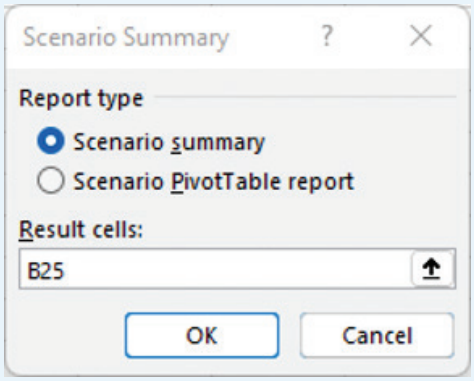


Figure 12.14 Scenario Summary Dialog Box



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.15 Scenario Summary for Middletown Amusement Park

Scenario Summary	Current Values:	Partly Cloudy	Rain	Sunny
Changing Cells:				
$B$6	3000	3000	1200	8000
$B$7	1600	1600	250	2400
$B$8	$15	$15	$10	$18
$B$9	$45	$45	$45	$57
$B$11	$33,000	$33,000	$27,000	$37,000
Result Cells:				
$B$25	$81,500	$81,500	-$6,625	$187,400

Notes: Current Values column represents values of changing cells at time Scenario Summary Report was created. Changing cells for each scenario are highlighted in gray.

The Scenario Summary report appears on a separate worksheet as shown in Figure 12.15. The summary includes the values currently in the spreadsheet, along with the specified scenarios. We see that the profit ranges from a low of −$6,625 on a rainy day to a high of $187,400 on a sunny day.

Notes + Comments

1. We emphasize the location of the reference to the desired output in a one-way versus a two-way data table. For a one-way data table, the reference to the output cell location is placed in the cell above and to the right of the column of input data so that it is in the cell just to the right of the label of the column of input data. For a two-way data table, the reference to the output cell location is placed above the column of input data and to the left of the row input data.
2. Notice that in Figures 12.5 and 12.7, the tables are formatted as currency. This must be done manually after the table is constructed using the options in the **Number** group under the **Home** tab in the Ribbon. It is also a good idea to label the rows and the columns of the table.
3. For very complex functions, Goal Seek might not converge to a stable solution. Trying several different initial values (the value in the cell referenced in the **By changing cell:** box when invoking **Goal Seek**) may help.
4. In Figure 12.13, we chose **Scenario summary** to generate the summary in Figure 12.14. Choosing **Scenario PivotTable report** will generate a pivot table with the relevant inputs and outputs.
5. Once all scenarios have been added to the **Scenario Manager** dialog box (Figure 12.11), there are several alternatives to choose. Scenarios can be edited via the **Edit...** button. The **Show** button allows you to look at the selected scenario settings by displaying the Scenario Values box for that scenario. The **Delete** button allows you to delete a scenario and the **Merge...** button allows you to merge scenarios from another worksheet with those of the current worksheet.

12.3 Some Useful Excel Functions for Modeling

In this section, we use several examples to introduce additional Excel functions that are useful in modeling decision problems. Many of these functions are used in spreadsheet models for simulation, optimization, and decision analysis.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

SUM and SUMPRODUCT

Two very useful functions are SUM and SUMPRODUCT. The SUM function adds up all of the numbers in a range of cells. The SUMPRODUCT function returns the sum of the products of elements in a set of arrays. As we shall see in Chapter 14, SUMPRODUCT is very useful for linear optimization models.

Let us illustrate the use of SUM and SUMPRODUCT by considering a transportation problem faced by Foster Generators. This problem involves the transportation of a product from three plants to four distribution centers. Foster Generators operates plants in Cleveland, Ohio; Bedford, Indiana; and York, Pennsylvania. Production capacities for the three plants over the next three-month planning period are known.

The firm distributes its generators through four regional distribution centers located in Boston, Massachusetts; Chicago, Illinois; St. Louis, Missouri; and Lexington, Kentucky. Foster has forecasted demand for the three-month period for each of the distribution centers. The per-unit shipping cost from each plant to each distribution center is also known. Management would like to determine how much of its products should be shipped from each plant to each distribution center.

A transportation analyst developed a what-if spreadsheet model to help Foster develop a plan for how to ship its generators from the plants to the distribution centers to minimize cost. Of course, capacity at the plants must not be exceeded, and forecasted demand must be satisfied at each of the four distribution centers. The what-if model is shown in Figure 12.16.

The parameters section is rows 2 through 10. Cells B5 through E7 contain the per-unit shipping cost from each origin (plant) to each destination (distribution center). For example, it costs $2.00 to ship one generator from Bedford to St. Louis. The plant capacities are given in cells F5 through F7, and the distribution center demands appear in cells B8 through E8.

The model is in rows 11 through 20. Trial values of shipment amounts from each plant to each distribution center appear in the shaded cells, B17 through E19. The total cost of shipping for this proposed plan is calculated in cell B13 using the SUMPRODUCT function. The general form of the SUMPRODUCT function is

=SUMPRODUCT(*array1*, *array2*)

The arrays used as arguments in the SUMPRODUCT function must be of the same dimension. For example, in the Foster Generator model, B5:E7 *is an array of three rows and four columns.* B17:E19 *is an array of the same dimensions.*

The function pairs each element of the first array with its counterpart in the second array, multiplies the elements of the pairs together, and adds the results. In cell B13, =SUMPRODUCT(B5:E7,B17:E19) pairs the per-unit cost of shipping for each origin-destination pair with the proposed shipping plan for that and adds their products:

$$B5*B17 + C5*C17 + D5*D17 + E5*E17 + B6*B18 + \cdots + E7*E19$$

In cells F17 through F19, the SUM function is used to add up the amounts shipped for each plant. The general form of the SUM function is

=SUM(*range*)

where *range* is a range of cells. For example, the function in cell F17 is =SUM(B17:E17), which adds the values in B17, C17, D17, and E17: $5000 + 0 + 0 + 0 = 5000$. The SUM function in cells B20 through E20 does the same for the amounts shipped to each distribution center.

We will revisit the Foster Generators problem in Chapter 14, where we discuss linear optimization models.

By comparing the amounts shipped from each plant to the capacity for that plant, we see that no plant violates its capacity. Likewise, by comparing the amounts shipped to each distribution center to the demand at that center, we see that all demands are met. The total shipping cost for the proposed plan is $54,500. Is this the lowest-cost plan? It is not clear.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.16 What-If Model for Foster Generators

	A	B	C	D	E	F	G
1	**Foster Generators**						
2	**Parameters**						
3	Shipping Cost/Unit			Destination			
4	Origin	Boston	Chcago	St. Louis	Lexington	Supply	
5	Cleveland	3	2	7	6	5000	
6	Bedford	6	5	2	3	6000	
7	York	2	5	4	5	2500	
8	Demand	6000	4000	2000	1500		
9							
10							
11	**Model**						
12							
13	Total Cost	=SUMPRODUCT(B5:E7,B17:E19)					
14							
15				Destination			
16	Origin	Boston	Chcago	St. Louis	Lexington	**Total**	
17	Cleveland	5000	0	0	0	=SUM(B17:E17)	
18	Bedford	1000	4000	1000	0	=SUM(B18:E18)	
19	York	0	0	1000	1500	=SUM(B19:E19)	
20	Total	=SUM(B17:B19)	=SUM(C17:C19)	=SUM(D17:D19)	=SUM(E17:E19)		
21							

MODEL *file*
foster

	A	B	C	D	E	F	G
1	**Foster Generators**						
2	**Parameters**						
3	Shipping Cost/Unit			Destination			
4	Origin	Boston	Chcago	St. Louis	exingtor	Supply	
5	Cleveland	$3.00	$2.00	$7.00	$6.00	5000	
6	Bedford	$6.00	$5.00	$2.00	$3.00	6000	
7	York	$2.00	$5.00	$4.00	$5.00	2500	
8	Demand	6000	4000	2000	1500		
9							
10							
11	**Model**						
12							
13	Total Cost	$54,500.00					
14							
15				Destination			
16	Origin	Boston	Chcago	St. Louis	exingtor	**Total**	
17	Cleveland	5000	0	0	0	5000	
18	Bedford	1000	4000	1000	0	6000	
19	York	0	0	1000	1500	2500	
20	Total	6000	4000	2000	1500		
21							

IF and COUNTIF

Gambrell Manufacturing produces car stereos. Stereos are composed of a variety of components that the company must carry in inventory to keep production running smoothly. However, because inventory can be a costly investment, Gambrell generally

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

likes to keep its components inventory to a minimum. To help monitor and control its inventory, Gambrell uses an inventory policy known as an *order-up-to policy*.

The order-up-to policy is as follows. Whenever the inventory on hand drops below a certain level, enough units are ordered to return the inventory to that predetermined level. If the current number of units in inventory, denoted by H, drops below M units, enough inventory is ordered to get the level back up to M units. M is called the *order-up-to point*. Stated mathematically, if Q is the amount we order, then

$$Q = M - H$$

An inventory model for Gambrell Manufacturing appears in Figure 12.17. In the upper half of the worksheet, the component ID number, inventory on hand (H), order-up-to point (M), and cost per unit are given for each of four components. Also given in this sheet is the fixed cost per order. The fixed cost is interpreted as follows: Each time a component is ordered, it costs Gambrell \$120 to process the order. The fixed cost of \$120 is incurred whenever an order is placed, regardless of how many units are ordered.

The model portion of the worksheet calculates the order quantity for each component. For example, for component 570, $M = 100$ and $H = 5$, so $Q = M - H = 100 - 5 = 95$. For component 741, $M = 70$ and $H = 70$ and no units are ordered because the on-hand inventory of 70 units is equal to the order-up-to point of 70. The calculations are similar for the other two components.

Depending on the number of units ordered, Gambrell receives a discount on the cost per unit. If 50 or more units are ordered, there is a quantity discount of 10% on every unit purchased. For example, for component 741, the cost per unit is \$4.50, and 95 units are ordered. Because 95 exceeds the 50-unit requirement, there is a 10% discount, and the cost per unit is reduced to \$4.50 − 0.1(\$4.50) = \$4.50 − \$0.45 = \$4.05. Not including the fixed cost, the cost of goods purchased is then \$4.05(95) = \$384.75.

The Excel functions used to perform these calculations are shown in Figure 12.17 (for clarity, we show formulas for only the first three columns). The IF function is used to calculate the purchase cost of goods for each component in row 17. The general form of the IF function is

=IF(*condition*, *result if condition is true*, *result if condition is false*)

For example, in cell B17 we have =IF(B16 >= \$B\$10, \$B\$11*B6, B6)*B16. This statement says that if the order quantity (cell B16) is greater than or equal to minimum amount required for a discount (cell B10), then the cost per unit is B11*B6 (there is a 10% discount, so the cost is 90% of the original cost); otherwise, there is no discount, and the cost per unit is the amount given in cell B6. The cost per unit computed by the IF function is then multiplied by the order quantity (B16) to obtain the total purchase cost of component 570. The purchase cost of goods for the other components are computed in a like manner.

The total cost in cell B23 is the sum of the total fixed ordering costs (B21) and the total cost of goods (B22). Because we place three orders (one each for components 570, 578, and 755), the fixed cost of the orders is 3*120 = \$360.

We introduce the COUNTIF *function in Chapter 2.*

The COUNTIF function in cell B19 is used to count how many times we order. In particular, it counts the number of components having a positive order quantity. The general form of the COUNTIF function is

=COUNTIF(*range*, *condition*)

The *range* is the range to search for the *condition*. The condition is the test to be counted when satisfied. In the Gambrell model in Figure 12.17, cell B19 counts the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.17 Gambrell Manufacturing Component Ordering Model

	A	B	C
1	**Gambrell Manufacturing**		
2	**Parameters**		
3	Component ID	**570**	**578**
4	Inventory On-Hand	5	30
5	Order Up to Point	100	55
6	Cost per unit	4.5	12.5
7			
8	Fixed Cost per Order	120	
9			
10	Minimum Order Size for Discount	50	
11	Discounted to	0.9	
12			
13	**Model**		
14			
15	Component ID	**=B3**	**=C3**
16	Order Quantity	=B5-B4	=C5-C4
17	Cost of Goods	=IF(B16 >= $B$10, $B$11*B6,B6)*B16	=IF(C16 >= $B$10, $B$11*C6,C6)*C16
18			
19	Total Number of Orders	=COUNTIF(B16:E16,">0")	
20			
21	Total Fixed Costs	=B19*B8	
22	Total Cost of Goods	=SUM(B17:E17)	
23	Total Cost	=SUM(B21:B22)	
24			

	A	B	C	D	E
1	**Gambrell Manufacturing**				
2	**Parameters**				
3	Component ID	**570**	**578**	**741**	**755**
4	Inventory On-Hand	5	30	70	17
5	Order Up to Point	100	55	70	45
6	Cost per unit	$4.50	$12.50	$3.26	$4.15
7					
8	Fixed Cost per Order	$120			
9					
10	Minimum Order Size for Discount	50			
11	Discounted to	90%			
12					
13	**Model**				
14					
15	Component ID	**570**	**578**	**741**	**755**
16	Order Quantity	95	25	0	28
17	Cost of Goods	$384.75	$312.50	$0.00	$116.20
18					
19	Total Number of Orders	3			
20					
21	Total Fixed Costs	$360.00			
22	Total Cost of Goods	$813.45			
23	Total Cost	$1,173.45			
24					

Notice the use of absolute references to B10 and B11 in row 17. As discussed in Appendix A, this facilitates copying cell B17 to cells C17, D17, and E17.

MODEL *file*
gambrell

number of cells that are greater than zero in the range of cells B16:E16 via the syntax =COUNTIF(B16:E16, ">0"). Note that quotes are required for the condition with the COUNTIF function. In the model, because only cells B16, C16, and E16 are greater than zero, the COUNTIF function in cell B19 returns 3.

As we have seen, IF and COUNTIF are powerful functions that allow us to make calculations based on a condition holding (or not). There are other such conditional functions available in Excel. In a problem at the end of this chapter, we ask you to investigate one such function, the SUMIF function. Another conditional function that is extremely useful in modeling is the XLOOKUP function, which is illustrated with an example in the next section.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

XLOOKUP

The Excel function XLOOKUP is not available is some older versions of Excel, but the functions VLOOKUP and HLOOKUP are similar.

The director of sales at Granite Insurance needs to award bonuses to their sales force based on performance. There are 15 salespeople, each with their own territory. Based on the size and population of the territory, each salesperson has a sales target for the year.

The measure of performance for awarding bonuses is the percentage achieved above the sales target. Based on this metric, a salesperson is placed into one of five bonus bands and awarded bonus points. After all salespeople are placed in a band and awarded points, each is awarded a percentage of the bonus pool, based on the percentage of the total points awarded. The sales director has created a spreadsheet model to calculate the bonuses to be awarded. The spreadsheet model is shown in Figure 12.18 (note that we have hidden rows 19–28).

As shown in cell E3 in Figure 12.18, the bonus pool is $250,000 for this year. The bonus bands are in cells A7:C11. In this table, column A gives the lower limit of the bonus band, column B the upper limit, and column C the bonus points awarded to anyone in that bonus band. For example, salespeople who achieve 56% above their sales target would be awarded 15 bonus points.

As shown in Figure 12.18, the name and percentage above the target achieved for each salesperson appear below the bonus-band table in columns A and B. In column C, the XLOOKUP function is used to look in the bonus band table and automatically assign the number of bonus points to each salesperson.

The XLOOKUP function allows the user to pull a subset of data from a larger table of data based on some criterion. The general form of the XLOOKUP function is

= XLOOKUP(*lookup_value, lookup_array, return_array, [if not found], [match mode], [search_mode]*)

Required arguments:

lookup_value = The value to search for; it can be an actual value or a cell location containing the value
lookup_array = The range to search over for the *lookup_value*
return_array = The range from which to return a value based on the *lookup_value*

Optional arguments:

[if not found] = when a valid match is not found, return the text you supply in *[if not found]*
The default return if a valid match is not found and this option is missing is #N/A
[match mode] = 0 means search for an exact match, if none is found, return #N/A (this is the default)
= −1 means search for an exact match, if none is found return the next smaller item
= 1 means search for an exact match, if none is found, return the next larger item
[search_mode] = 1 means start the search at the first item in the list
= −1 means start the search at the last item in the list (a reverse search)
= 2 means perform a search that relies on *lookup_array* being sorted in ascending order. If not sorted, invalid results will be returned.
= −2 means perform a search that relies on *lookup_array* being sorted in descending order. If not sorted, invalid results will be returned.

If [match mode] *is set to −1 or 1, then the data should be sorted, and* [search_mode] *set to either 2 or −2.*

The XLOOKUP function used to assign a bonus points for salesperson Choi in cell C18 is as follows:

= XLOOKUP(B18,$A$7:$A$11,$C$7:$C$11,#N/A, −1,2)

This function uses the percentage above target sales from cell B18 and searches the range A7:A11. Because *[match mode]* = −1 and *[search_mode]* = 2, it searches

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.18 Granite Insurance Bonus Model

	A	B	C	D	E
1	**Granite Insurance Bonus Awards**				
2					
3	**Parameters**			Bonus Pool	250000
4					
5	Bonus Bands to be awarded for percentage above target sales.				
6	**Lower Limit**	**Upper Limit**	**Bonus Points**		
7	0	0.1	0		
8	0.11	0.5	10		
9	0.51	0.79	15		
10	0.8	0.99	25		
11	1	100	40		
12					
13	**Model**				
14	**Last Name**	**% Above Target Sales**	**Bonus Points**	**% of Pool**	**Bonus Amount**
15	Barth	0.83	=XLOOKUP(B15,$A$7:$A$11,$C$7:$C$11,#N/A,-1,2)	=C15/$C$30	=D15*$E$3
16	Benson	0	=XLOOKUP(B16,$A$7:$A$11,$C$7:$C$11,#N/A,-1,2)	=C16/$C$30	=D16*$E$3
17	Capel	1.18	=XLOOKUP(B17,$A$7:$A$11,$C$7:$C$11,#N/A,-1,2)	=C17/$C$30	=D17*$E$3
18	Choi	0.44	=XLOOKUP(B18,$A$7:$A$11,$C$7:$C$11,#N/A,-1,2)	=C18/$C$30	=D18*$E$3
29	Ruebush	0.85	=XLOOKUP(B29,$A$7:$A$11,$C$7:$C$11,#N/A,-1,2)	=C29/$C$30	=D29*$E$3
30		Total	=SUM(C15:C29)	=SUM(D15:D29)	=SUM(E15:E29)

MODEL *file*
granite

	A	B	C	D	E
1	**Granite Insurance Bonus Awards**				
2					
3	**Parameters**			Bonus Pool	$250,000
4					
5	Bonus Bands to be awarded for percentage above target sales.				
6	**Lower Limit**	**Upper Limit**	**Bonus Points**		
7	0%	10%	0		
8	11%	50%	10		
9	51%	79%	15		
10	80%	99%	25		
11	100%	10000%	40		
12					
13	**Model**				
14	**Last Name**	**% Above Target Sales**	**Bonus Points**	**% of Pool**	**Bonus Amount**
15	Barth	83%	25	8.5%	$21,186.44
16	Benson	0%	0	0.0%	$0.00
17	Capel	118%	40	13.6%	$33,898.31
18	Choi	44%	10	3.4%	$8,474.58
29	Ruebush	85%	25	8.5%	$21,186.44
30		Total	295	100%	$250,000.00

A7:A11 until it finds a number strictly greater than the value of B18. B18 is 44%, and the first value in the table in column A larger than 44% is in cell A9 (51%). It then backs up one row (to row 8). In other words, it finds the last value in the first column less than or equal to 44%. Because the *return_array* is set to C7:C11, it returns the element in row 8 (cell C8), which is 10 bonus points. In summary, XLOOKUP searches for what is in cell B18 until it finds a larger value and then backs up one row and returns the value in that row from the range C7:C11.

Once all salespeople are awarded bonus points based on XLOOKUP and the bonus-band table, the total number of bonus points awarded is given in cell C30 using the SUM function. Each person's bonus points as a percentage of the total awarded is calculated in column D, and in column E each person is awarded that percentage of the bonus pool. As a check, cells D30 and E30 give the total percentages and dollar amounts awarded.

Numerous mathematical, logical, and financial functions are available in Excel. In addition to those discussed here, we will introduce you to other functions, as needed, in examples and end-of-chapter problems. Having already discussed principles for building

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

good spreadsheet models and after having seen a variety of spreadsheet models, we turn now to how to audit Excel models to ensure model integrity.

12.4 Auditing Spreadsheet Models

Excel contains a variety of tools to assist you in the development and debugging of spreadsheet models. These tools are found in the **Formula Auditing** group of the **Formulas** tab, as shown in Figure 12.19. Let us review each of the tools available in this group.

Trace Precedents and Dependents

After selecting cells, the Trace Precedents button creates arrows pointing to the selected cell from cells that are part of the formula in that cell. The Trace Dependents button, on the other hand, shows arrows pointing from the selected cell to cells that depend on the selected cell. Both of the tools are excellent for quickly ascertaining how parts of a model are linked.

An example of Trace Precedents is shown in Figure 12.20. Here we have opened the Foster Generators Excel file, selected cell B13, and clicked the **Trace Precedents** button in the **Formula Auditing** group. Recall that the cost in cell B13 is calculated as the SUMPRODUCT of the per-unit shipping cost and units shipped. In Figure 12.20, to show this relationship, arrows are drawn to these areas of the spreadsheet to cell B13. These arrows may be removed by clicking on the **Remove Arrows** button in the **Formula Auditing** group.

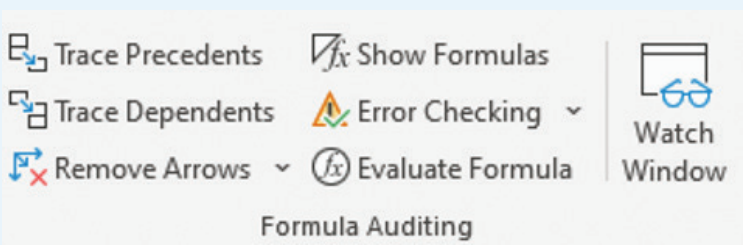
An example of Trace Dependents is shown in Figure 12.21. We have selected cell E18, the units shipped from Bedford to Lexington, and clicked on the **Trace Dependents** button in the **Formula Auditing** group. As shown in Figure 12.21, units shipped from Bedford to Lexington impacts the cost function in cell B13, the total units shipped from Bedford given in cell F18, as well as the total units shipped to Lexington in cell E20. These arrows may be removed by clicking on the **Remove Arrows** button in the **Formula Auditing** group.

Trace Precedents and Trace Dependents can highlight errors in copying and formula construction by showing that incorrect sections of the worksheet are referenced.

Show Formulas

The Show Formulas button does exactly that. To see the formulas in a worksheet, simply click on any cell in the worksheet and then click on **Show Formulas**. You will see the formulas residing in that worksheet. To revert to hiding the formulas, click again on the **Show Formulas** button. As we have already seen in our examples in this chapter, the use of Show Formulas allows you to inspect each formula in detail in its cell location.

Figure 12.19 The Formula Auditing Group



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.20 Trace Precedents for Foster Generator

	A	B	C	D	E	F	G
1	**Foster Generators**						
2	**Parameters**						
3	Shipping Cost/Unit			Destination			
4	Origin	Boston	Chcago	St. Louis	Lexington	Supply	
5	Cleveland	$3.00	$2.00	$7.00	$6.00	5000	
6	Bedford	$6.00	$5.00	$2.00	$3.00	6000	
7	York	$2.00	$5.00	$4.00	$5.00	2500	
8	Demand	6000	4000	2000	1500		
9							
10							
11	**Model**						
12							
13	Total Cost	$54,500.00					
14							
15				Destination			
16	Origin	Boston	Chcago	St. Louis	Lexington	**Total**	
17	Cleveland	5000	0	0	0	5000	
18	Bedford	1000	4000	1000	0	6000	
19	York	0	0	1000	1500	2500	
20	Total	6000	4000	2000	1500		
21							

Figure 12.21 Trace Dependents for the Foster Generators Model

	A	B	C	D	E	F	G
1	**Foster Generators**						
2	**Parameters**						
3	Shipping Cost/Unit			Destination			
4	Origin	Boston	Chcago	St. Louis	Lexington	Supply	
5	Cleveland	$3.00	$2.00	$7.00	$6.00	5000	
6	Bedford	$6.00	$5.00	$2.00	$3.00	6000	
7	York	$2.00	$5.00	$4.00	$5.00	2500	
8	Demand	6000	4000	2000	1500		
9							
10							
11	**Model**						
12							
13	Total Cost	$54,500.00					
14							
15				Destination			
16	Origin	Boston	Chcago	St. Louis	Lexington	**Total**	
17	Cleveland	5000	0	0	0	5000	
18	Bedford	1000	4000	1000	0	6000	
19	York	0	0	1000	1500	2500	
20	Total	6000	4000	2000	1500		
21							

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Evaluate Formulas

The **Evaluate Formula** button allows you to investigate the calculations of a cell in great detail. As an example, let us investigate cell B17 of the Gambrell Manufacturing model (Figure 12.17). Recall that we are calculating cost of goods based on whether there is a quantity discount. We follow these steps:

MODEL file
gambrell

Step 1. Select cell B17
Step 2. Click the **Formulas** tab in the Ribbon
Step 3. Click the **Evaluate Formula** button (fx Evaluate Formula) in the **Formula Auditing** group
Step 4. When the **Evaluate Formula** dialog box appears (Figure 12.22), click the **Evaluate** button
Step 5. Repeat Step 4 until the formula has been completely evaluated
Step 6. Click **Close**

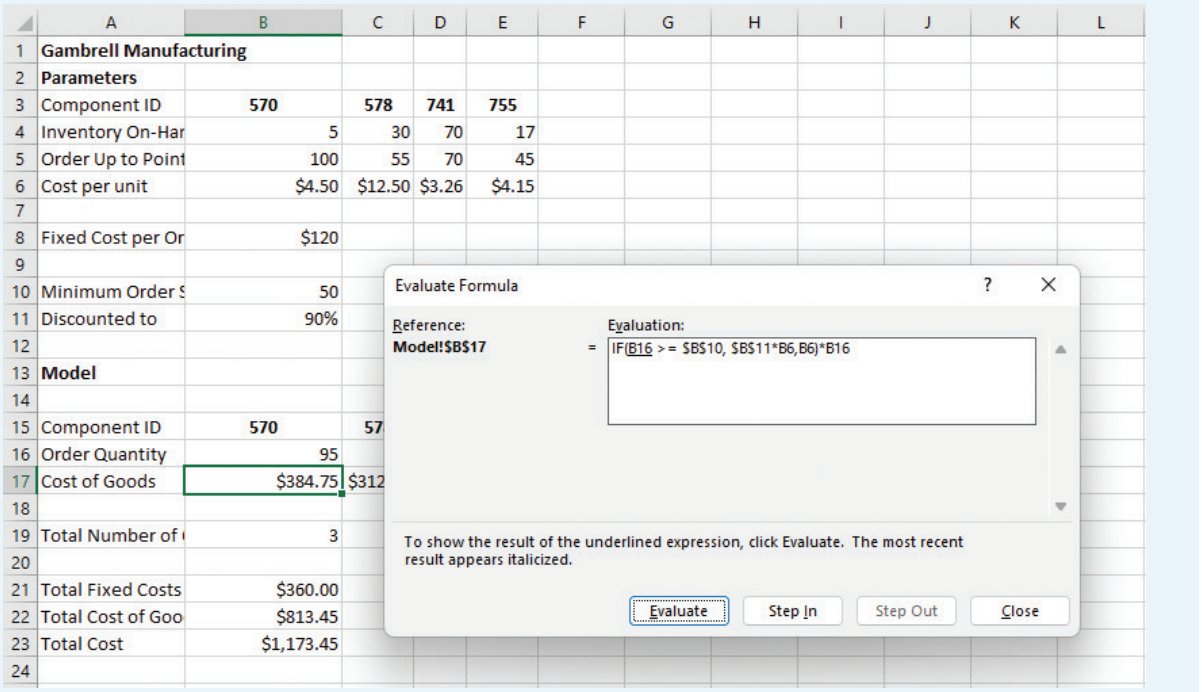
Figure 12.23 shows the **Evaluate Formula** dialog box for cell B17 in the Gambrell Manufacturing spreadsheet model after four clicks of the **Evaluate** button.

The Evaluate Formula tool provides an excellent means of identifying the exact location of an error in a formula.

Error Checking

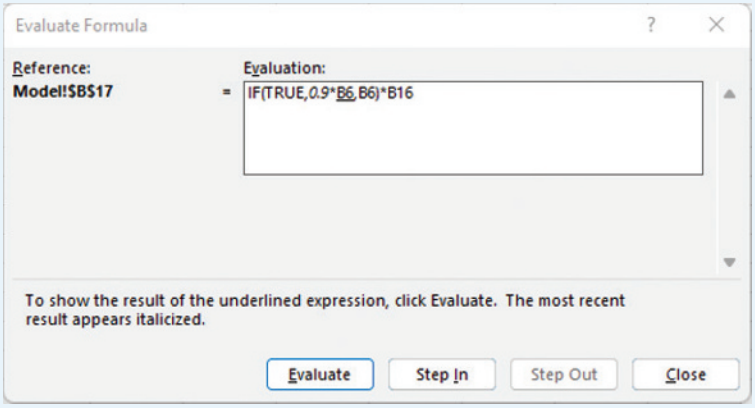
The **Error Checking** button provides an automatic means of checking for mathematical errors within formulas of a worksheet. Clicking on the **Error Checking** button causes Excel to check every formula in the sheet for calculation errors. If an error is found, the **Error Checking** dialog box appears. An example for a hypothetical

Figure 12.22 The Evaluate Formula Dialog Box for Gambrell Manufacturing



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.23 The Evaluate Formula Dialog Box for Gambrell Manufacturing Cell B17 After Four Clicks of the Evaluate Button



division by zero error is shown in Figure 12.24. From this box, the formula can be edited, the calculation steps can be observed (as in the previous section on Evaluate Formulas), or help can be obtained through the Excel help function. The Error Checking procedure is particularly helpful for large models where not all cells of the model are visible.

Watch Window

The **Watch Window**, located in the **Formula Auditing** group, allows the user to observe the values of cells included in the Watch Window box list. This is useful for large models when not all of the model is observable on the screen or when multiple worksheets are used. The user can monitor how the listed cells change with a change in the model without searching through the worksheet or changing from one worksheet to another.

A Watch Window for the Gambrell Manufacturing model is shown in Figure 12.25. The following steps were used to add cell B17 to the watch list:

Step 1. Click the **Formulas** tab in the Ribbon
Step 2. Click **Watch Window** in the **Formula Auditing** group to display the **Watch Window**
Step 3. Click **Add Watch…**
Step 4. Select the cell you would like to add to the watch list (in this case B17)

As shown in Figure 12.25, the list gives the workbook name, worksheet name, cell name (if used), cell location, cell value, and cell formula. To delete a cell from the watch list, click on the entry from the list, and then click on the **Delete Watch** button that appears in the upper part of the **Watch Window**.

The Watch Window, as shown in Figure 12.25, allows us to monitor the value of B17 as we make changes elsewhere in the worksheet. Furthermore, if we had other worksheets in this workbook, we could monitor changes to B17 of the worksheet even from these other worksheets. The Watch Window is observable regardless of where we are in any worksheet of a workbook.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 12.24 The Error Checking Dialog Box for a Division by Zero Error

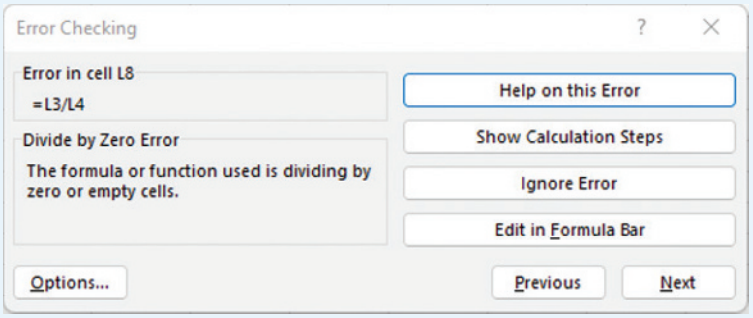
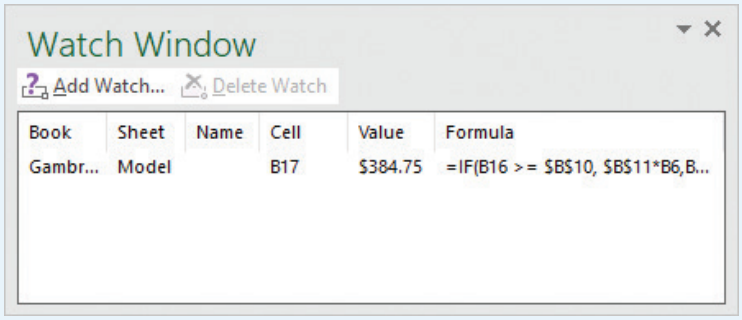


Figure 12.25 The Watch Window for Cell B17 of the Gambrell Manufacturing Model



12.5 Predictive and Prescriptive Spreadsheet Models

Two key phenomena that make decision making difficult are uncertainty and an overwhelming number of choices. Spreadsheet what-if models, as we have discussed thus far in this chapter, are descriptive models. Given formulas and data to populate the formulas, calculations are made based on the formulas. However, basic what-if spreadsheet models can be extended to help deal with uncertainty or the many alternatives a decision maker may face.

As we have seen in previous chapters, predictive models can be estimated from data in spreadsheets using tools provided in Excel. For example, the Excel Regression tool and other Data Analysis tools such as Exponential Smoothing and Moving Average allow us to develop predictive models based on data in the spreadsheet. These predictive models can help us deal with uncertainty by giving estimates for unknown events/quantities that serve as inputs to the decision-making process.

Monte Carlo simulation is discussed in detail in Chapter 13.

Another important extension of what-if models that helps us deal with uncertainty is simulation. Monte Carlo simulation automates a manual what-if model by replacing the manual evaluation of various values of input parameters with the random generation of values for these uncertain inputs. By generating the values of the uncertain inputs with the likelihoods that the analyst believes these values may occur in the future, the simulation model allows the analyst to conduct many experiments to quantify how the uncertainty

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

in the inputs affects the uncertainty in the output measures. Excel has built-in probability functions that allow us to simulate values of uncertain inputs.

In Chapters 14, 15, and 16, we discuss optimization models for decision making and how to use Excel Solver.

To deal with the other complicating factor of decision making, namely an overwhelming number of alternatives, optimization models can be used to help make smart decisions. Optimization models are prescriptive models, characterized by having an objective to be maximized or minimized and usually have constraints that limit the options available to the decision maker. Because they yield a course of action to follow, optimization models are one type of prescriptive analytics. Excel includes a special tool called Solver that solves optimization models. Excel Solver is used to extend a what-if model to find an optimal (or best) course of action that maximizes or minimizes an objective while satisfying the constraints of the decision problem.

Summary

In this chapter, we discussed the principles of building good spreadsheet models, several what-if analysis tools, some useful Excel functions, and how to audit spreadsheet models. What-if spreadsheet models are important and popular analysis tools in and of themselves, but as we shall see in later chapters, they also serve as the basis for optimization and simulation models.

We discussed how to use influence diagrams to structure a problem. Influence diagrams can serve as a guide to developing a mathematical model and implementing the model in a spreadsheet. We discussed the importance of separating the parameters from the model because it leads to simpler analysis and minimizes the risk of creating an error in a formula. In most cases, cell formulas should use cell references in their arguments rather than being "hardwired" with values. We also discussed the use of proper formatting and color to enhance the ease of use and understanding of a spreadsheet model.

We used examples to illustrate how Excel What-If Analysis tools Data Tables, Goal Seek, and Scenario Manager can be used to perform detailed and efficient what-if analysis. We also discussed a number of Excel functions that are useful for business analytics. We discussed Excel Formula Auditing tools that may be used to debug and monitor spreadsheet models to ensure that they are error-free and accurate. We ended the chapter with a brief discussion of predictive and prescriptive spreadsheet models.

Glossary

Data Table An Excel tool that quantifies the impact of changing the value of a specific input on an output of interest.
Decision variable A model input the decision maker can control.
Goal Seek An Excel tool that allows the user to determine the value for an input cell that will cause the value of a related output cell to equal some specified value, called the *goal*.
Influence diagram A visual representation that shows which entities influence others in a model.
Make-versus-buy decision A decision often faced by companies that have to decide whether they should manufacture a product or outsource its production to another firm.
One-way data table An Excel Data Table that summarizes a single input's impact on the output of interest.
Parameters In a what-if model, the uncontrollable model input.
Scenario manager An Excel tool that quantifies the impact of changing multiple inputs on one or more outputs of interest.
Two-way data table An Excel Data Table that summarizes two inputs' impact on the output of interest.
What-if model A model designed to study the impact of changes in model inputs on model outputs.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Problems

1. **Profit Model for Electronics Company.** Cox Electric makes electronic components and has estimated the following for a new design of one of its products:

   Fixed cost = $10,000
   Material cost per unit = $0.15
   Labor cost per unit = $0.10
   Revenue per unit = $0.65

   DATA*file* coxelectric

   These data are given in the file *coxelectric*. Note that fixed cost is incurred regardless of the amount produced. Per-unit material and labor cost together make up the variable cost per unit. Assuming that Cox Electric sells all that it produces, profit is calculated by subtracting the fixed cost and total variable cost from total revenue. **LO 1, 2**
   a. Build an influence diagram that illustrates how to calculate profit.
   b. Using mathematical notation similar to that used for Nowlin Plastics, give a mathematical model for calculating profit.
   c. Implement your model from part (b) in Excel using the principles of good spreadsheet design.
   d. If Cox Electric makes 12,000 units of the new product, what is the resulting profit?

2. **Breakeven Volume.** Use the spreadsheet model constructed to answer Problem 1 to answer this problem. **LO 3, 4, 5, 7**
   a. Construct a one-way data table with production volume as the column input and profit as the output. Breakeven occurs when profit goes from a negative to a positive value; that is, breakeven is when total revenue = total cost, yielding a profit of zero. Vary production volume from 0 to 100,000 in increments of 10,000. In which interval of production volume does breakeven occur?
   b. Use Goal Seek to find the exact breakeven point. Assign **Set cell:** equal to the location of profit, **To value:** = 0, and **By changing cell:** equal to the location of the production volume in your model.

3. **E-book Profit/Loss.** Eastman Publishing Company is considering publishing an electronic textbook about spreadsheet applications for business. The fixed cost of manuscript preparation, textbook design, and web site construction is estimated to be $160,000. Variable processing costs are estimated to be $6 per book. The publisher plans to sell single-user access to the book for $46. **LO 1, 2, 3**
   a. Create an influence diagram to calculate the profit/loss.
   b. Build a spreadsheet model to calculate the profit/loss for a given demand. What profit can be anticipated with a demand of 3,500 copies?
   c. Use a data table to vary demand from 1,000 to 6,000 in increments of 200 to assess the sensitivity of profit to demand.

4. **E-book Profit/Loss (revisited).** Consider the Profit/Loss problem presented in problem 3. **LO 5, 6**
   a. Use Goal Seek to determine the access price per copy that the publisher must charge to break even with a demand of 3,500 copies.
   b. Consider the following scenarios:

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Variable Cost/ Book	$6	$8	$12	$10	$11
Access Price	$46	$50	$40	$50	$60
Demand	2,500	1,000	6,000	5,000	2,000

   For each of these scenarios, the fixed cost remains $160,000. Use Scenario Manager to generate a summary report that gives the profit for each of these scenarios. Which scenario yields the highest profit? Which scenario yields the lowest profit?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

5. **Cash Flow Planning for a Microbrewery.** Over the Rhine Brewery (ORB) is conducting a cash flow analysis for the coming year. The first quarter of the calendar year is typically the slowest time for ORB. For each month of the first quarter (January, February, and March), the estimate for the sales of beverages and food is $60,000 per month. Using the first quarter as the base period, ORB has analyzed historical data and has established quarterly multipliers for each of the other three quarters of the year to estimate monthly sales in those quarters. The multiplier for months in the first quarter is 1.0 because this is the base period. The other multipliers are 1.5 for the second quarter, 3.2 for the third quarter, and 2.8 for the fourth quarter. So, for example, the estimate for beverage and food sales in the month of November, since November is in the fourth quarter, would be (2.8)($60,000) = $168,000. ORB also sells merchandise such as hats and t-shirts. In a given month, revenue from merchandise is estimated to be about 8% of the sales from beverage and food. Material cost is 23% of total sales revenue and labor cost is 45% of total sales revenue. ORB pays $3,000 per month for rent and $1,300 per month for utilities. ORB borrowed money through a small-business loan during its startup and currently makes $1,200 monthly loan payments (70% of the payment is interest and 30% is principal). The interest of the loan is tax deductible, but the principal portion of the loan is not. ORB's tax rate is 23.6%. Assume ORB begins with a zero cash balance at the beginning of the year. **LO 1, 2, 5**
   a. Build an influence diagram for ORB's monthly cash flow problem for the coming year.
   b. Use the influence diagram from part (a) to construct a spreadsheet model. What is the ending total cash balance at the end of the year?
   c. ORB would like to increase the sales of merchandise. Use Goal Seek to find the percentage of food and beverage sales that revenue from merchandise would have to be to provide an end-of-year cash balance of $400,000.
6. **Breakeven Analysis for a Symposium.** The University of Cincinnati Center for Business Analytics is an outreach center that collaborates with industry partners on applied research and continuing education in business analytics. One of the programs offered by the center is a quarterly Business Intelligence Symposium. Each symposium features three speakers on the real-world use of analytics. Each corporate member of the center (there are currently 10) receives five free seats to each symposium. Nonmembers wishing to attend must pay $75 per person. Each attendee receives breakfast, lunch, and free parking. The following are the costs incurred for putting on this event: **LO 1, 2, 4, 5**

Rental cost for the auditorium	$150
Registration processing	$8.50 per person
Speaker costs	3@$800 = $2,400
Continental breakfast	$4.00 per person
Lunch	$7.00 per person
Parking	$5.00 per person

   a. Create an influence diagram of how to calculate profit/loss.
   b. Use your influence diagram from part (a) to build a spreadsheet model that calculates a profit or loss based on the number of nonmember registrants.
   c. Use Goal Seek to find the number of nonmember registrants that will make the event break even.
7. **Scenario Analysis for Profitability of a Symposium.** Consider again the scenario described in Problem 6. **LO 2, 3, 6, 7**
   a. The Center for Business Analytics is considering a refund policy for no-shows. No refund would be given for members who do not attend, but nonmembers who do not attend will be refunded 50% of the price. Extend the model you developed in Problem 4 for the Business Intelligence Symposium to account for the fact that, historically, 25% of members who registered do not show and 10% of registered

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

nonmembers do not attend. The center pays the caterer for breakfast and lunch based on the number of registrants (not the number of attendees). However, the center pays for parking only for those who attend. What is the profit if each corporate member registers their full allotment of tickets and 127 nonmembers register?

b. Use a two-way data table to show how profit changes as a function of number of registered nonmembers and the no-show percentage of nonmembers. Vary the number of nonmember registrants from 80 to 160 in increments of 5 and the percentage of nonmember no-shows from 10% to 30% in increments of 2%.

c. Consider three scenarios:

	Base Case	Worst Case	Best Case
% of Members Who Do Not Show	25.0%	50%	15%
% of Nonmembers Who Do Not Show	10.0%	30%	5%
Number of Nonmember Registrants	130	100	150

All other inputs are the same as in part (a). Use Scenario Manager to generate a summary report that gives the profit for each of these three scenarios. What is the highest profit? What is the lowest profit?

8. **E-book Profit/Loss (Revisited).** Consider again Problem 3. Through a series of web-based experiments, Eastman has created a predictive model that estimates demand as a function of price. The predictive model is demand $= 4{,}000 - 6p$, where p is the price of the e-book. **LO 2, 3, 5, 7**
   a. Update your spreadsheet model constructed for Problem 3 to take into account this demand function.
   b. Use Goal Seek to calculate the price that results in breakeven.
   c. Use a data table that varies price from $50 to $400 in increments of $25 to find the price that maximizes profit.

9. **Retirement Planning.** Lindsay is 25 years old and has a new job in web development. Lindsay wants to make sure that they are financially sound in 30 years. So, Lindsay plans to invest the same amount into a retirement account at the end of every year for the next 30 years. Note that because Lindsay invests at the end of the year, there is no interest earned on the contribution for the year in which Lindsay contributes. **LO 2, 3**
   a. Construct a spreadsheet model that will calculate Lindsay's retirement savings given a fixed annual amount saved and a fixed annual interest rate of return.
   b. Develop a two-way table for annual investment amounts of $5,000 to $20,000 in increments of $1,000 and for returns of 0% to 12% in increments of 1%.

10. **Retirement Planning with Net Present Value.** Consider again Lindsay's investment in Problem 9. The real value of Lindsay's account after 30 years of investing will depend on inflation over that period. In the Excel function =NPV(*rate*, *value1, value2, . . .*), *rate* is called the discount rate, and *value 1*, *value 2*, etc. are incomes (positive) or expenditures (negative) over equal periods of time. Update your model from Problem 9 using the NPV function to get the net present value of Lindsay's retirement fund. Construct a data table that shows the net present value of Lindsay's retirement fund for various levels of return and inflation (discount rate). Use a data table to vary the return from 0% to 12% in increments of 1% and the discount rate from 0% to 4% in increments of 1% to show the impact on the net present value. *Hint:* Calculate the total amount added to the account each year, and discount that stream of payments using the NPV function. **LO 2, 3, 7**

11. **Powerball Winner.** Californian Vinh Nguyen won the Powerball Jackpot and was the sole winner. The jackpot was $228,467,735. Nguyen had the option of taking a lump sum of $134 million or an annuity that paid 30 annual payments such that the payments increased by 5% each year and added up to the total jackpot of $228,467,735. Nguyen opted for the annuity. The tax rate is 39% and inflation is expected to be 3%. **LO 2, 5, 7**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Construct a spreadsheet model and use Goal Seek to calculate the payment schedule, the amount Nguyen would receive each year for the annuity. What is the amount of the thirtieth payment?
b. Use the Excel NPV function to determine the net present value (NPV) of Nguyen's 30 payments after taxes. Assume the first annual payment occurs immediately. *Hint*: The Excel function NPV is discussed in problem 10.
c. Compare the NPV of the annuity to the lump sum after taxes (assuming lump sum payment occurs immediately)
d. Use Goal Seek to determine the inflation rate that make the NPV of the annuity equal to the net value of the lump sum after taxes.

12. **Net Discounted Cash Flow.** Goal Kick Sports (GKS) is a retail chain that sells youth and adult soccer equipment. The GKS financial planning group has developed a spreadsheet model to calculate the net discounted cash flow of the first five years of operations for a new store. This model is used to assess new locations under consideration for expansion. **LO 6, 7, 8**

MODEL *file*
goalkick

a. Use Excel's Formula Auditing tools to audit this model and correct any errors found.
b. Once you are comfortable that the model is correct, use Scenario Manager to generate a Scenario Summary report that gives Total Discounted Cash Flow for the following scenarios:

	Scenario			
	1	2	3	4
Tax Rate	33%	25%	33%	25%
Inflation Rate	1%	2%	4%	3%
Annual Growth of Sales	20%	15%	10%	12%

What is the range of values for the Total Discounted Cash Flow for these scenarios?

13. **Analyzing a Supply Chain.** Newton Manufacturing produces scientific calculators. The models are N350, N450, and N900. Newton has planned its distribution of these products around eight customer zones: Brazil, China, France, Malaysia, U.S. Northeast, U.S. Southeast, U.S. Midwest, and U.S. West. Data for the current quarter (volume to be shipped in thousands of units) for each product and each customer zone are given in the file *newton*. Newton would like to know the total number of units going to each customer zone and also the total units of each product shipped. There are several ways to get this information from the data set. One way is to use the SUMIF function.

DATA *file*
newton

The SUMIF function extends the SUM function by allowing the user to add the values of cells meeting a logical condition. The general form of the function is

$$=\text{SUMIF}(\textit{test range}, \textit{condition}, \textit{range to be summed})$$

The *test range* is an area to search to test the *condition*, and the *range to be summed* is the position of the data to be summed. So, for example, using the file *newton*, we use the following function to get the total units sent to Malaysia:

$$=\text{SUMIF}(\text{A3:A26}, \text{A3}, \text{C3:C26})$$

Cell A3 contains the text "Malaysia"; A3:A26 is the range of customer zones; and C3:C26 are the volumes for each product for these customer zones. The SUMIF looks for matches of "Malaysia" in column A and, if a match is found, adds the volume to the total. Use the SUMIF function to get total volume by each zone and total volume by each product. **LO 2, 7**

14. **Auditing a Transportation Model.** Consider the transportation model in the file *williamson*, which is very similar to the Foster Generators model discussed in this chapter. Williamson produces a single product and has plants in Atlanta, Lexington, Chicago, and Salt Lake City and warehouses in Portland, St. Paul, Las Vegas, Tucson, and Cleveland. Each plant has a capacity, and each warehouse has a demand. Williamson

MODEL *file*
williamson

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

would like to find a low-cost shipping plan. Mr. Williamson has reviewed the results and notices right away that the total cost is way out of line. Use the **Formula Auditing** tool under the **Formulas** tab in Excel to find any errors in this model. Correct the errors. *Hint:* The model contains two errors. Be sure to check every formula. **LO 8**

15. **Calculating Course Grades.** Professor Rao would like to accurately calculate the grades for the 58 students in their Operations Planning and Scheduling class (OM 455). Rao has thus far constructed a spreadsheet, part of which follows: **LO 2, 7**

MODEL *file*
om455

	A	B	C	D	E
1	OM 455				
2	Section 001				
3	Course Grading Scale Based on Course Average:				
4		Lower	Upper	Course	
5		Limit	Limit	Grade	
6		0	59	F	
7		60	69	D	
8		70	79	C	
9		80	89	B	
10		90	100	A	
11					
12		Midterm	Final	Course	Course
13	Last Name	Score	Score	Average	Grade
14	Alt	70	56	63.0	
15	Amini	95	91	93.0	
16	Amoako	82	80	81.0	
17	Apland	45	78	61.5	
18	Bachman	68	45	56.5	
19	Corder	91	98	94.5	
20	Desi	87	74	80.5	
21	Dransman	60	80	70.0	
22	Duffuor	80	93	86.5	
23	Finkel	97	98	97.5	
24	Foster	90	91	90.5	

a. The Course Average is calculated by weighting the Midterm Score and Final Score 50% each. Use the XLOOKUP function with the table shown to generate the Course Grade for each student in cells E14 through E24.
b. Use the COUNTIF function to determine the number of students receiving each letter grade.

16. **Revenue Model with Quantity Discounts.** Richardson Ski Racing (RSR) sells equipment needed for downhill ski racing. One of RSR's products is fencing used on downhill courses. The fence product comes in 150-foot rolls and sells for $215 per roll. However, RSR offers quantity discounts. The following table shows the price per roll depending on order size:

DATA *file*
rsr

Quantity Ordered		
From	To	Price per Roll ($)
1	50	215
51	100	195
101	200	175
201	∞	155

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The file *rsr* contains 172 orders that have arrived for the coming six weeks. LO 2, 7

a. Use the XLOOKUP function with the preceding pricing table to determine the total revenue from these orders.
b. Use the COUNTIF function to determine the number of orders in each price bin.

17. **Currency Exchange Rates.** Companies that compete globally must pay attention to exchange rates between currencies. The exchange rate between two currencies depends on a variety of factors including each country's gross domestic product, rate of economic growth, interest rate, and rate of unemployment. Consider the case of one U.S. dollar and one Euro with the following exchange rate: 1.00 U.S. dollar = 0.95 Euros. If a company in the United States sells its product in Germany for €50 (50 Euros), what is that revenue worth in U.S. dollars? To convert €50 to U.S. dollars at the exchange rate of 1.00 U.S. dollar = 0.95 Euros, we solve the following equation for x:

$$\frac{x}{€50} = \frac{\$1}{€0.95}$$

DATA *file* globalrevenue

This results in $x = 50 \div 0.95 = \$52.63$. So, €50 is equal to 52.63 U.S. dollars. The file *globalrevenue* contains 299 payments received by a company in the United States, paid in various currencies. Also given in the file are the exchange rates for one U.S. dollar for 53 different currencies. For each payment, the order number, the amount of the payment and the currency of the payment are given. LO 2, 7

a. Using the XLOOKUP function to find the relevant exchange rate, convert each payment to U.S. dollars. Round the U.S. dollars to two decimal places using the ROUND function. Note that to place the rounded value of cell D2 in cell E2, type into cell E2 the following function =ROUND(D2,2). What was the U.S. dollar value of Order #1039?
b. Use the SUM function and the rounded values to calculate the total dollar amount received from the 299 orders in U.S. dollars. What is the total amount received in U.S. dollars?

18. **State Sales Tax for Online Purchases.** For online purchases, the location of the buyer and the location of the business (seller) determine if state sales tax must be collected at the time of the sale. The rule for collecting sales tax is that if the business has a physical presence in a state, they must collect applicable sales taxes from online customers who reside in that state. Breckenridge Fabrics sells fabric online and has distribution centers in California, Florida, New Jersey, Ohio, and Texas. The file *salestax* contains sales data on 279 purchases. For each purchase the following are provided: Order Number, Buyer State, Supplier State (state of the distribution center that supplied the goods) and the Sale Amount (in dollars). Also, in the file *salestax* is a state tax table that gives the applicable sales tax rate for each of the 50 states and the District of Columbia. LO 2, 7

DATA *file* salestax

a. For each of the 279 purchases, use the XLOOKUP and IF functions to determine if sales tax must be collected and if so, how much. Round the sales tax due to two decimal places using the ROUND function. What is the total amount of state sales taxes collected from these 279 purchases?
b. What proportion of these purchases had to pay sales tax?
c. What is the total amount of these purchases (including any sales tax)?

19. **European Financial Options.** A put option in finance allows you to sell a share of stock at a given price in the future. There are different types of put options. A European put option allows you to sell a share of stock at a given price, called the exercise price, at a particular point in time after the purchase of the option. For example, suppose you purchase a six-month European put option for a share of stock with an exercise price of $26. If six months later, the stock price per share is $26 or more, the option has no value. If in six months the stock price is lower than $26 per share, then you can purchase the stock and immediately sell it at the higher exercise price of $26. If the price per share in six months is $22.50, you can purchase a share of the stock for

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$22.50 and then use the put option to immediately sell the share for $26. Your profit would be the difference, $26 − $22.50 = $3.50 per share, less the cost of the option. If you paid $1.00 per put option, then your profit would be $3.50 − $1.00 = $2.50 per share. **LO 2, 3**

a. Build a model to calculate the profit of this European put option.
b. Construct a data table that shows the profit per share for a share price in six months between $10 and $30 per share in increments of $1.00.

20. **European Financial Options (Revisited).** Consider again Problem 19. The point of purchasing a European option is to limit the risk of a decrease in the per-share price of the stock. Suppose you purchased 200 shares of the stock at $28 per share and 75 six-month European put options with an exercise price of $26. Each put option costs $1. **LO 2, 3**
    a. Using data tables, construct a model that shows the value of the portfolio with options and without options for a share price in six months between $15 and $35 per share in increments of $1.00.
    b. Discuss the value of the portfolio with and without the European put options.

21. **Revenue with Substitutable Products.** The Camera Shop sells two popular models of digital single lens reflex (DSLR) cameras. The sales of these products are not independent; if the price of one increases, the sales of the other increases. In economics, these two camera models are called *substitutable products*. The store wishes to establish a pricing policy to maximize revenue from these products. A study of price and sales data shows the following relationships between the quantity sold (N) and price (P) of each model. **LO 2, 3**

$$N_A = 195 - 0.6P_A + 0.25P_B$$
$$N_B = 301 + 0.08P_A - 0.5P_B$$

    a. Construct a model for the total revenue and implement it on a spreadsheet.
    b. Develop a two-way data table to estimate the optimal prices for each product in order to maximize the total revenue. Vary each price from $250 to $500 in increments of $10.

22. **Refinancing a Mortgage.** A few years back, Dave and Jana bought a new home. They borrowed $230,415 at an annual fixed rate of 5.49% (15-year term) with monthly payments of $1,881.46. They just made their 25th payment, and the current balance on the loan is $208,555.87.

    Interest rates are at an all-time low, and Dave and Jana are thinking of refinancing to a new 15-year fixed loan. Their bank has made the following offer: 15-year term, 3.0%, plus out-of-pocket costs of $2,937. The out-of-pocket costs must be paid in full at the time of refinancing.

    Build a spreadsheet model to evaluate this offer. The Excel function

$$=\text{PMT}(rate, nper, pv, fv, type)$$

    calculates the payment for a loan based on constant payments and a constant interest rate. The arguments of this function are as follows:

$rate$ = the interest rate for the loan
$nper$ = the total number of payments
pv = present value (the amount borrowed)
fv = future value [the desired cash balance after the last payment (usually 0)]
$type$ = payment type (0 = end of period, 1 = beginning of the period)

    For example, for Dave and Jana's original loan, there will be 180 payments (12*15 = 180), so we would use =PMT(0.0549/12, 180, 230415, 0, 0) = $1,881.46. Note that because payments are made monthly, the annual interest rate must be expressed as a monthly rate. Also, for payment calculations, we assume that the payment is made at the end of the month.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The savings from refinancing occur over time, and therefore need to be discounted back to current dollars. The formula for converting K dollars saved t months from now to current dollars is

$$\frac{K}{(1+r)^{t-1}}$$

where r is the monthly inflation rate. Assume that $r = 0.002$ and that Dave and Jana make their payment at the end of each month.

Use your model to calculate the savings in current dollars associated with the refinanced loan versus staying with the original loan. **LO 2, 7**

23. **Mortgage Prepayment.** Consider again the mortgage refinance problem in Problem 22. Assume that Dave and Jana have accepted the refinance offer of a 15-year loan at 3% interest rate with out-of-pocket expenses of $2,937. Recall that they are borrowing $208,555.87. Assume that there is no prepayment penalty, so that any amount over the required payment is applied to the principal. Construct a model so that you can use Goal Seek to determine the monthly payment that will allow Dave and Jana to pay off the loan in 12 years. Do the same for 10 and 11 years. Which option for prepayment, if any, would you choose and why? *Hint:* Break each monthly payment up into interest and principal [the amount that is deducted from the balance owed]. Recall that the monthly interest that is charged is the monthly loan rate multiplied by the remaining loan balance. **LO 2, 5**

24. **Assigning Customers to Distribution Centers.** Floyd's Bumpers has distribution centers in Lafayette, Indiana; Charlotte, North Carolina; Los Angeles, California; Dallas, Texas; and Pittsburgh, Pennsylvania. Each distribution center carries all products sold. Floyd's customers are auto repair shops and larger auto parts retail stores. You are asked to perform an analysis of the customer assignments to determine which of Floyd's customers should be assigned to each distribution center. The rule for assigning customers to distribution centers is simple: A customer should be assigned to the closest center. The file *floyds* contains the distance from each of Floyd's 1,029 customers to each of the five distribution centers. Your task is to build a list that tells which distribution center should serve each customer. The following functions will be helpful:

DATA *file*
floyds

$$=\text{MIN}(array)$$

The MIN function returns the smallest value in a set of numbers. For example, if the range A1:A3 contains the values 6, 25, and 38, then the formula =MIN(A1:A3) returns the number 6, because it is the smallest of the three numbers.

$$=\text{MATCH}(lookup_value, lookup_array, match_type)$$

The MATCH function searches for a specified item in a range of cells and returns the relative position of that item in the range. The *lookup_value* is the value to match, the *lookup_array* is the range of search, and *match_type* indicates the type of match (use 0 for an exact match).

For example, if the range A1:A3 contains the values 6, 25, and 38, then the formula =MATCH(25,A1:A3,0) returns the number 2, because 25 is the second item in the range.

$$=\text{INDEX}(array, column_num)$$

The INDEX function returns the value of an element in a position of an array. For example, if the range A1:A3 contains the values 6, 25, and 38, then the formula =INDEX(A1: A3, 2) = 25, because 25 is the value in the second position of the array A1:A3. *Hint:* Create three new columns. In the first column, use the MIN function to calculate the minimum distance for the customer in that row. In the second column use the MATCH function to find the position of the minimum distance. In the third column, use the position in the previous column with the INDEX function referencing the row of distribution center names to find the name of the distribution center that should service that customer. **LO 2, 7**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



25. **Transportation Costs.** Refer to Problem 24. Floyd's Bumpers pays a transportation company to ship its product in full truckloads to its customers. Therefore, the cost for shipping is a function of the distance traveled and a fuel surcharge (also on a per-mile basis). The cost per mile is $2.42, and the fuel surcharge is $0.56 per mile. The file *floydsmay* contains data for shipments for the month of May (each record is simply the customer zip code for a given truckload shipment) as well as the distance table from the distribution centers to each customer. Use the MATCH and INDEX functions to retrieve the distance traveled for each shipment, and calculate the charge for each shipment. What is the total amount that Floyd's Bumpers spends on these May shipments? *Hint:* The INDEX function may be used with a two-dimensional array: =INDEX(*array*, *row_num*, *column_num*), where *array* is a matrix, *row_num* is the row number, and *column_num* is the column position of the desired element of the matrix. **LO 2, 7**

26. **Discount Price versus Zero-Percent Financing.** An auto dealership is advertising that a new car with a sticker price of $35,208 is on sale for $25,995 if payment is made in full, or it can be financed at 0% interest for 72 months with a monthly payment of $489. Note that 72 payments × $489 per payment = $35,208, which is the sticker price of the car. By allowing you to pay for the car in a series of payments (starting one month from now) rather than $25,995 now, the dealer is effectively loaning you $25,995. If you choose the 0% financing option, what is the effective interest rate that the auto dealership is earning on your loan? *Hint:* Discount the payments back to current dollars [see Problem 22 for a discussion of discounting], and use Goal Seek to find the discount rate that makes the net present value of the payments = $25,995. **LO 2, 5**

Case Problem: Retirement Plan

Tim is 37 years old and would like to establish a retirement plan. Develop a spreadsheet model that could be used to assist Tim with retirement planning. Your model should include the following input parameters:

Tim's current age = 37 years
Tim's current total retirement savings = $259,000
Annual rate of return on retirement savings = 4%
Tim's current annual salary = $145,000
Tim's expected annual percentage increase in salary = 2%
Tim's percentage of annual salary contributed to retirement = 6%
Tim's expected age of retirement = 65
Tim's expected annual expenses after retirement (current dollars) = $90,000
Rate of return on retirement savings after retirement = 3%
Income tax rate postretirement = 15%

Assume that Tim's employer contributes 6% of Tim's salary to their retirement fund. Tim can make an additional annual contribution to their retirement fund before taxes (tax free) up to a contribution of $16,000. Assume that Tim contributes $6,000 per year. Also, assume an inflation rate of 2%. **LO 2, 3**

Managerial Report

Your spreadsheet model should provide the accumulated savings at the onset of retirement as well as the age at which funds will be depleted (given assumptions on the input parameters).

Include a data table in your spreadsheet model to demonstrate the sensitivity of the age at which funds will be depleted to the retirement age and additional pre-tax contributions. Similarly, consider other factors you think might be important.

Develop a report for Tim outlining the factors that will have the greatest impact on Tim's retirement.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 13

Monte Carlo Simulation

Contents

Analytics in Action: Evaluating Financial Risk for Loan Providers

13.1 Risk Analysis for Sanotronics LLC
- Base-Case Scenario
- Worst-Case Scenario
- Best-Case Scenario
- Sanotronics Spreadsheet Model
- Use of Probability Distributions to Represent Random Variables
- Generating Values for Random Variables with Excel
- Executing Simulation Trials with Excel
- Measuring and Analyzing Simulation Output

13.2 Inventory Policy Analysis for Promus Corp
- Spreadsheet Model for Promus
- Generating Values for Promus Corp's Demand
- Executing Simulation Trials and Analyzing Output

13.3 Simulation Modeling for Land Shark Inc.
- Spreadsheet Model for Land Shark
- Generating Values for Land Shark's Random Variables
- Executing Simulation Trials and Analyzing Output
- Generating Bid Amounts with Fitted Distributions

13.4 Simulation with Dependent Random Variables
- Spreadsheet Model for Press Teag Worldwide

13.5 Simulation Considerations
- Verification and Validation
- Advantages and Disadvantages of Using Simulation

Summary 715
Glossary 716
Problems 717
Appendix 13.1: Common Probability Distributions for Simulation 734

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Distinguish between the base-case, worst-case, and best-case scenarios in a decision-making problem with uncertainty.

LO 2 Construct and interpret a simulation model that implements a continuous probability distribution to model an uncertain quantity.

LO 3 Construct and interpret a simulation model that implements a discrete probability distribution to model an uncertain quantity.

LO 4 Construct and interpret a simulation model that directly samples values from a set of observations to model an uncertain quantity.

LO 5 Identify situations in simulation modeling when there is dependency between uncertain quantities and appropriately model this dependency.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

Evaluating Financial Risk for Loan Providers*

The financial crisis of 2008 led to nearly 9 million lost jobs in 2008 and 2009 in the United States, approximately 6% of the overall workforce. The net worth of U.S. households declined by nearly $13 trillion, and the U.S. stock market fell by almost 50%. One of the major factors leading to this financial crisis was an increase in subprime mortgage defaults. A subprime mortgage is a loan granted to individuals who do not qualify for conventional loans because they have poor credit scores. These types of mortgages are exposed to considerable financial risk that was not completely understood at the time.

While the financial markets and overall economy in the United States largely recovered from the 2008 financial crisis, many companies still face similar financial risk from large pools of loans. Even today, large banks in the United States often own more than a million mortgages. Loan servicers in the United States service up to 10 million mortgages. These types of companies face substantial financial risk in the forms of delinquency risk, meaning customers who fail to pay back their loans on time, as well as prepayment risk which is driven by customers who choose to pay off their loans early resulting in reduced cash flows for the loan providers.

Complicated equations are available for some forms of delinquency and prepayment risk for large pools of loans. However, these equations are sensitive to input parameters and often rely on strict assumptions regarding the distributions of delinquent payments and prepayments. Therefore, Monte Carlo simulation is a common tool used to quantify delinquency and prepayment risk.

In one particular study, historical data were used from 10 million subprime mortgages from across the United States across 36,000 different zip codes as well as a separate set of 16 million mortgages insured by Freddie Mac, a U.S. government sponsored agency that buys mortgages and packages them into mortgage-backed securities for investors. The data sets include loan information such as credit scores, whether or not the mortgagee is a first-time home buyer, type of mortgage, debt-to-income values for the mortgagee, and zip code of the mortgage. The data sets also include information on prepayment and default events. This historical information can then be used to create analytical models to predict delinquency and prepayments.

The predictive models then feed into a Monte Carlo simulation model that can be used to estimate future delinquency and prepayment risks based on these many different input parameters. The input parameters can be varied to estimate the sensitivity of delinquency and prepayment risks to input parameters such as the location of the mortgages in a loan pool and the number of first-time home buyers in a loan pool. The goal of such models is to help companies better estimate the financial risks present in their pools of loans so that they can adequately plan for future events.

*Partially based on information from J. Sirignano and K. Giesecke, "Risk Analysis for Large Pools of Loans," *Management Science*, vol. 65, no. 1 (2019): 107–121.

Monte Carlo simulation originated during World War II as part of the Manhattan Project to develop nuclear weapons. "Monte Carlo" was selected as the code name for the classified method in reference to the famous Monte Carlo casino in Monaco and the uncertainties inherent in gambling.

Uncertainty pervades decision making in business, government, and our personal lives. This chapter introduces the use of **Monte Carlo simulation** to evaluate the impact of uncertainty on a decision. Simulation models have been successfully used in a variety of disciplines. Financial applications include investment planning, project selection, and option pricing. Marketing applications include new product development and the timing of market entry for a product. Management applications include project management, inventory ordering (especially important for seasonal products), capacity planning, and revenue management (prominent in the airline, hotel, and car rental industries). Monte Carlo simulation models are used by healthcare companies to estimate treatment costs and evaluate public health risks. Sports teams and sports analysts use Monte Carlo simulation to forecast performance. In each of these applications, uncertain quantities complicate the decision process.

As we will demonstrate, a spreadsheet simulation analysis requires a model foundation of logical formulas that correctly express the relationships between parameters and decisions to generate outputs of interest. For example, a clothing retailer may implement a simple spreadsheet model to compute profit given values for the number of ski jackets ordered from the manufacturer and the number of ski jackets demanded by customers. A simulation analysis extends this model by replacing the single value used for ski jacket demand with a

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

probability distribution of possible values of ski jacket demand. A probability distribution of ski jacket demand represents not only the range of possible values but also the relative likelihood of various levels of demand.

In general, to evaluate a decision with a Monte Carlo simulation, an analyst identifies parameters that are not known with a high degree of certainty and treats these parameters as random variables. A **random variable** is a quantity whose value is uncertain and whose possible values are described by a probability distribution. In a simulation model, different values for one or more random variables are randomly generated from specified probability distributions. Then, the simulation model uses the randomly generated values of the random variables and the relationships between parameters and decisions to compute the corresponding values of an output. Specifically, a simulation experiment produces a *distribution* of output values that correspond to the randomly generated values of the uncertain input variables. This probability distribution of the output values describes the range of possible outcomes, as well as the relative likelihood of each outcome. After reviewing the simulation results, the analyst is often able to make decision recommendations for the **controllable inputs** that address not only the *average* output but also the *variability* of the output.

In this chapter, we construct spreadsheet simulation models using only native Excel functionality. As we will show, practical simulation models for real-world problems can be executed in native Excel. However, there are many simulation software products that provide sophisticated simulation modeling features and automate the generation of outputs such as charts and summary statistics. Some of these software packages can be installed as Excel add-ins, including @RISK, Crystal Ball, and Analytic Solver.

13.1 Risk Analysis for Sanotronics LLC

When making a decision in the presence of uncertainty, the decision maker should be interested not only in the average, or expected, outcome, but also in information regarding the range of possible outcomes. In particular, decision makers are interested in **risk analysis**, that is, quantifying the likelihood and magnitude of an undesirable outcome. In this section, we show how to perform a risk analysis study for a medical device company.

Sanotronics LLC is a start-up company that manufactures medical devices for use in hospital clinics. Inspired by experiences with family members who have battled cancer, Sanotronics's founders have developed a prototype for a new device that limits health care workers' exposure to chemotherapy treatments while they are preparing, administering, and disposing of these hazardous medications. The new device features an innovative design and has the potential to capture a substantial share of the market.

Sanotronics would like an analysis of the first-year profit potential for the device. Because of Sanotronics's tight cash flow situation, management is particularly concerned about the potential for a loss. Sanotronics has identified the key parameters in determining first-year profit: selling price per unit (p), first-year administrative and advertising costs (c_a), direct labor cost per unit (c_i), parts cost per unit (c_p), and first-year demand (d). After conducting market research and a financial analysis, Sanotronics estimates with a high level of certainty that the device's selling price will be \$249 per unit and that the first-year administrative and advertising costs will total \$1,000,000.

Sanotronics is not certain about the values for the cost of direct labor, the cost of parts, and the first-year demand. At this stage of the planning process, Sanotronics's base estimates of these inputs are \$45 per unit for the direct labor cost, \$90 per unit for the parts cost, and 15,000 units for the first-year demand. We begin our risk analysis by considering a small set of what-if scenarios.

Base-Case Scenario

Sanotronics's first-year profit is computed as follows:

$$\text{Profit} = (p - c_i - c_p) \times d - c_a \tag{13.1}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Recall that Sanotronics is certain of a selling price of $249 per unit, and administrative and advertising costs total $1,000,000. Substituting these values into equation (13.1) yields

$$\text{Profit} = (249 - c_i - c_p) \times d - 1{,}000{,}000 \quad \textbf{(13.2)}$$

Sanotronics's base-case estimates of the direct labor cost per unit, the parts cost per unit, and first-year demand are $45, $90, and 15,000 units, respectively. These values constitute the **base-case scenario** for Sanotronics. Substituting these values into equation (13.2) yields the following profit projection:

$$\text{Profit} = (249 - 45 - 90)(15{,}000) - 1{,}000{,}000 = 710{,}000$$

Thus, the base-case scenario leads to an anticipated profit of $710,000.

Although the base-case scenario looks appealing, Sanotronics is aware that the values of direct labor cost per unit, parts cost per unit, and first-year demand are uncertain, so the base-case scenario may not occur. To help Sanotronics gauge the impact of the uncertainty, the company may consider performing a what-if analysis. A **what-if analysis** involves considering alternative values for the random variables (direct labor cost, parts cost, and first-year demand) and computing the resulting value for the output (profit).

What-if analysis is introduced in Chapter 12.

Sanotronics is interested in what happens if the estimates of the direct labor cost per unit, parts cost per unit, and first-year demand do not turn out to be as expected under the base-case scenario. For instance, suppose that Sanotronics believes that direct labor costs could range from $43 to $47 per unit, parts cost could range from $80 to $100 per unit, and first-year demand could range from 0 to 30,000 units. Using these ranges, what-if analysis can be used to evaluate a **worst-case scenario** and a **best-case scenario**.

Worst-Case Scenario

The worst-case scenario for the direct labor cost is $47 (the highest value), the worst-case scenario for the parts cost is $100 (the highest value), and the worst-case scenario for demand is 0 units (the lowest value). Substituting these values into equation (13.2) leads to the following profit projection:

$$\text{Profit} = (249 - 47 - 100)(0) - 1{,}000{,}000 = -1{,}000{,}000$$

So, the worst-case scenario leads to a projected *loss* of $1,000,000.

Best-Case Scenario

The best-case value for the direct labor cost is $43 (the lowest value), for the parts cost it is $80 (the lowest value), and for demand it is 30,000 units (the highest value). Substituting these values into equation (14.2) leads to the following profit projection:

$$\text{Profit} = (249 - 43 - 80)(30{,}000) - 1{,}000{,}000 = 2{,}780{,}000$$

So the best-case scenario leads to a projected *profit* of $2,780,000.

At this point, the what-if analysis provides the conclusion that profits may range from a loss of $1,000,000 to a profit of $2,780,000 with a base-case profit of $710,000. Although the base-case profit of $710,000 is possible, the what-if analysis indicates that either a substantial loss or a substantial profit is also possible. Sanotronics can repeat this what-if analysis for other scenarios. However, simple what-if analyses do not indicate the likelihood of the various profit or loss values. In particular, we do not know anything about the probability of a loss. To conduct a more thorough evaluation of risk by obtaining insight on the potential magnitude and probability of undesirable outcomes, we now turn to developing a spreadsheet simulation model.

Sanotronics Spreadsheet Model

The first step in constructing a spreadsheet simulation model is to express the relationship between the inputs and the outputs with appropriate formula logic. Figure 13.1 provides the formula and value views for the Sanotronics spreadsheet. Data on selling price per

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.1 What-If Spreadsheet Model for Sanotronics

	A	B
1	**Sanotronics**	
2		
3	**Parameters**	
4	Selling Price per Unit	249
5	Administrative & Advertising Cost	1000000
6	Direct Labor Cost Per Unit	45
7	Parts Cost Per Unit	90
8	Demand	15000
9		
10	**Model**	
11	Profit	=((B4-B6-B7)*B8)-B5
12	Profit ($ million)	=B11/1000000

	A	B
1	**Sanotronics**	
2		
3	**Parameters**	
4	Selling Price per Unit	$249.00
5	Administrative & Advertising Cost	$1,000,000
6	Direct Labor Cost Per Unit	$45.00
7	Parts Cost Per Unit	$90.00
8	Demand	15,000
9		
10	**Model**	
11	Profit	$702,215
12	Profit ($ million)	$0.702

unit, administrative and advertising cost, direct labor cost per unit, parts cost per unit, and demand are in cells B4 to B8. The profit calculation, corresponding to equation (13.1), is expressed in cell B11 using appropriate cell references and formula logic. In cell B12, profit is converted into being measured in millions of dollars (instead of dollars as in B11). This conversion is not strictly necessary, but as we construct the simulation model, it will reduce visual clutter of the output. For the values shown in Figure 13.1, the spreadsheet model computes profit for the base-case scenario. By changing one or more values for the input parameters, the spreadsheet model can be used to conduct a manual what-if analysis (e.g., the best-case and worst-case scenarios).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Use of Probability Distributions to Represent Random Variables

Using the what-if approach to risk analysis, we manually select values for the random variables (direct labor cost per unit, parts cost per unit, and first-year demand), and then compute the resulting profit. Instead of manually selecting the values for the random variables, a Monte Carlo simulation randomly generates values for the random variables according to specified probability distributions that appropriately reflect the uncertainty in these respective quantities. A probability distribution describes the possible values of a random variable and the relative likelihood of the random variable taking on these values. The analyst can use historical data and knowledge of the random variable (range, mean, mode, and standard deviation) to specify the probability distribution for a random variable. As we describe in the following paragraphs, Sanotronics researched the direct labor cost per unit, the parts cost per unit, and first-year demand to identify the respective probability distributions for these three random variables.

Probability distributions are discussed in more detail in Chapter 5.

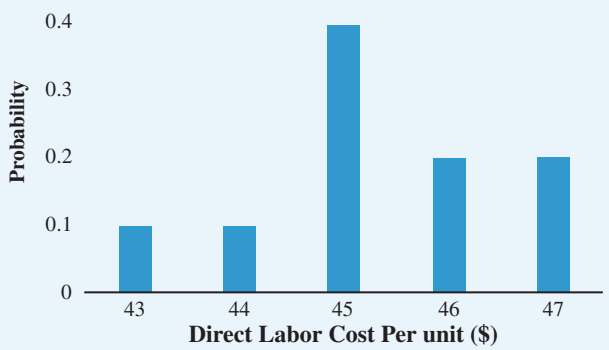
Based on recent wage rates and estimated processing requirements of the device, Sanotronics believes that the direct labor cost will range from \$43 to \$47 per unit and is described by the discrete probability distribution shown in Figure 13.2. We see that there is a 0.1 probability that the direct labor cost will be \$43 per unit, a 0.1 probability that the direct labor cost will be \$44 per unit, and so on. The highest probability, 0.4, is associated with a direct labor cost of \$45 per unit. Because we have assumed that the direct labor cost per unit is best described by a **discrete probability distribution**, the direct labor cost per unit can take on *only* the values of \$43, \$44, \$45, \$46, or \$47.

One advantage of simulation is that the analyst can adjust the probability distributions of the random variables to determine the impact of the assumptions about the shape of the uncertainty on the output measures.

Sanotronics is relatively unsure of the parts cost because it depends on many factors, including the general economy, the overall demand for parts, and the pricing policy of Sanotronics's parts suppliers. Sanotronics is confident that the parts cost will be between \$80 and \$100 per unit but is unsure as to whether any particular values between \$80 and \$100 are more likely than others. Therefore, Sanotronics decides to describe the uncertainty in parts cost with a uniform probability distribution, as shown in Figure 13.3. Costs per unit between \$80 and \$100 are equally likely. A uniform probability distribution is an example of a **continuous probability distribution**, which means that the parts cost can take on *any* value between \$80 and \$100.

Based on sales of comparable medical devices, Sanotronics believes that first-year demand is described by the normal probability distribution shown in Figure 13.4. The mean μ of first-year demand is 15,000 units. The standard deviation σ of 4,500 units describes the variability in the first-year demand. The normal probability distribution is a continuous probability distribution in which any value is possible, but values extremely larger or smaller than the mean are increasingly unlikely.

Figure 13.2 Probability Distribution for Direct Labor Cost per Unit



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.3 Uniform Probability Distribution for Parts Cost per Unit

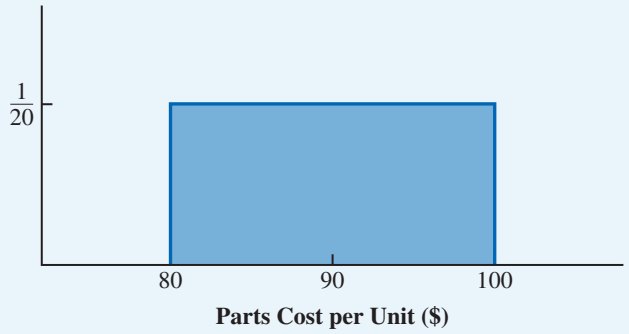
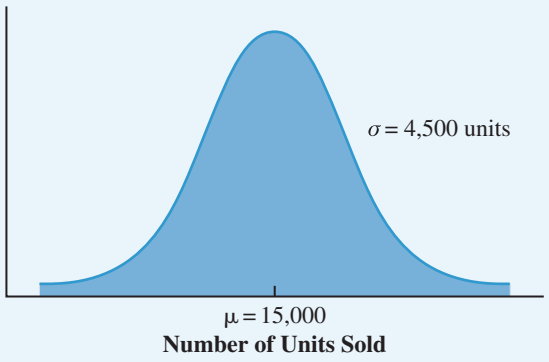


Figure 13.4 Normal Probability Distribution for First-Year Demand



Generating Values for Random Variables with Excel

A set of values for the random variables and the values of the associated output measures for a single scenario in a simulation model is called a **trial**. In the Sanotronics model, a trial consists of values for direct labor cost per unit, parts cost per unit, first-year demand, and the resulting profit. To execute a simulation for the Sanotronics example, we must generate values for the three random variables and compute the resulting profit. Then we generate another trial, compute a second value for profit, and so on. We continue this process until we are satisfied that enough trials have been conducted to describe the probability distribution for profit. Put simply, simulation is the process of generating values of random variables and computing the corresponding output measures.

In the Sanotronics model, respective values must be generated for the random variables corresponding to direct labor cost per unit, the parts cost per unit, and the first-year demand. To illustrate how to generate these values, we need to introduce the concept of computer-generated random numbers.

Computer-generated random numbers[1] are randomly selected numbers from 0 up to, but not including, 1; this interval is denoted by [0, 1). All values of the computer-generated

[1]Computer-generated random numbers are formally called pseudorandom numbers because they are generated through the use of mathematical formulas and are therefore not technically random. The difference between random numbers and pseudorandom numbers is primarily philosophical, and we use the term *random numbers* even when they are generated by a computer.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

random numbers are equally likely and so the values are uniformly distributed over the interval from 0 to 1. Computer-generated random numbers can be obtained using built-in functions available in computer simulation packages and spreadsheets. For example, placing the formula =RAND() in a cell of an Excel worksheet will result in a random number between 0 and 1 being placed into that cell.

Let us show how random numbers can be used to generate values corresponding to the probability distributions for the random variables in the Sanotronics example. We begin by showing how to generate a value for the direct labor cost per unit. The approach described is applicable for generating values from any discrete probability distribution.

Table 13.1 illustrates the process of partitioning the interval from 0 to 1 into subintervals so that the probability of generating a random number in a subinterval is equal to the probability of the corresponding direct labor cost. The interval of random numbers from 0 up to but not including 0.1, denoted by [0, 0.1), is associated with a direct labor cost of $43; the interval of random numbers from 0.1 up to but not including 0.2, denoted by [0.1, 0.2), is associated with a direct labor cost of $44, and so on. With this assignment of random number intervals to the possible values of the direct labor cost, the probability of generating a random number in any interval is equal to the probability of obtaining the corresponding value for the direct labor cost. Thus, to select a value for the direct labor cost, we generate a random number between 0 and 1 using the RAND function in Excel. If the random number is at least 0.0 but less than 0.1, we set the direct labor cost equal to $43. If the random number is at least 0.1 but less than 0.2, we set the direct labor cost equal to $44, and so on.

Each trial of the simulation requires a value for the direct labor cost. Suppose that on the first trial the random number is 0.9109. From Table 13.1, because 0.9109 is in the interval [0.8, 1.0), the corresponding simulated value for the direct labor cost would be $47 per unit. Suppose that on the second trial the random number is 0.2841. From Table 13.1, the simulated value for the direct labor cost would be $45 per unit.

Each trial in the simulation also requires a value of the parts cost and first-year demand. Let us now turn to the issue of generating values for the parts cost. The probability distribution for the parts cost per unit is the uniform distribution shown in Figure 13.3. Because this random variable has a different probability distribution than direct labor cost, we use random numbers in a slightly different way to generate simulated values for parts cost. To generate a value for a random variable characterized by a continuous uniform distribution, the following Excel formula is used:

$$\text{Value of uniform random variable} = \text{lower bound} + (\text{upper bound} - \text{lower bound}) \times \text{RAND}() \qquad \textbf{(13.3)}$$

Table 13.1 Random Number Intervals for Generating Value of Direct Labor Cost per Unit

Direct Labor Cost per Unit ($)	Probability	Interval of Random Numbers	Cumulative Probability
43	0.1	[0.0, 0.1)	0.1
44	0.1	[0.1, 0.2)	0.2
45	0.4	[0.2, 0.6)	0.6
46	0.2	[0.6, 0.8)	0.8
47	0.2	[0.8, 1.0)	1.0

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For Sanotronics, the parts cost per unit is a uniformly distributed random variable with a lower bound of \$80 and an upper bound of \$100. Applying equation (13.3) leads to the following formula for generating the parts cost:

$$\text{Parts cost} = 80 + 20 \times \text{RAND}() \quad \textbf{(13.4)}$$

By closely examining equation (13.4), we can understand how it uses random numbers to generate uniformly distributed values for parts cost. The first term of equation (13.4) is 80 because Sanotronics is assuming that the parts cost will never drop below \$80 per unit. Because RAND() returns a value between 0 and 1, the second term, $20 \times$ RAND(), corresponds to how much more than the lower bound the simulated value of parts cost is. Because RAND() is equally likely to be any value between 0 and 1, the simulated value for the parts cost is equally likely to be between the lower bound ($80 + 0 = 80$) and the upper bound ($80 + 20 = 100$). For example, suppose that a random number of 0.4576 is generated by the RAND function. As illustrated by Figure 13.5, the value for the parts cost would be

$$\text{Parts cost} = 80 + 20 \times 0.4576 = 80 + 9.15 = 89.15 \text{ per unit}$$

Suppose that a random number of 0.5842 is generated on the next trial. The value for the parts cost would be

$$\text{Parts cost} = 80 + 20 \times 0.5842 = 80 + 11.68 = 91.68 \text{ per unit}$$

With appropriate choices of the lower and upper bounds, equation (13.3) can be used to generate values for any continuous uniform probability distribution.

Lastly, we need a procedure for generating the first-year demand from computer-generated random numbers. Because first-year demand is normally distributed with a mean of 15,000 units and a standard deviation of 4,500 units (see Figure 13.4), we need a procedure for generating random values from this normal probability distribution.

Once again we will use random numbers between 0 and 1 to simulate values for first-year demand. To generate a value for a random variable characterized by a normal distribution with a specified mean and standard deviation, the following Excel formula is used:

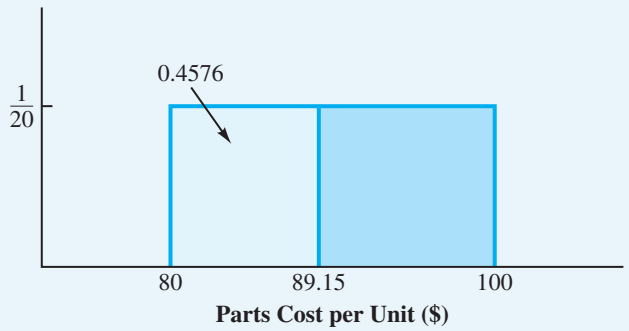
Equation (13.5) can be used to generate values for any normal probability distribution by changing the values specified for the mean and standard deviation, respectively.

$$\text{Value of normal random variable} = \text{NORM.INV(RAND(), mean, standard deviation)} \quad \textbf{(13.5)}$$

For Sanotronics, first-year demand is a normally distributed random variable with a mean of 15,000 and a standard deviation of 4,500. Applying equation (13.5) leads to the following formula for generating the first-year demand:

$$\text{Demand} = \text{NORM.INV(RAND(), 15000, 4500)} \quad \textbf{(13.6)}$$

Figure 13.5 Generation of Value for Parts Cost per Unit Corresponding to Random Number 0.4576



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

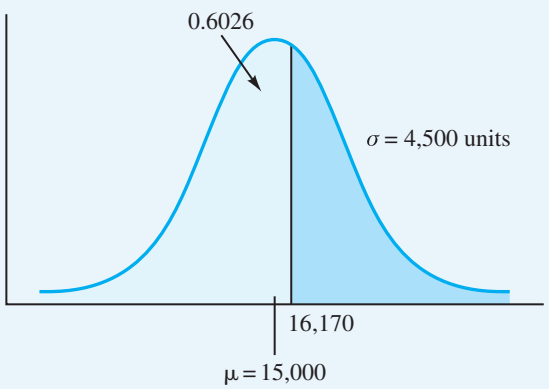
Suppose that the random number of 0.6026 is produced by the RAND function; applying equation (13.6) then results in Demand =NORM.INV(0.6026, 15000, 4500) = 16,170 units. To understand how equation (13.6) uses random numbers to generate normally distributed values for first-year demand, observe from Figure 13.6 that 60.26% of the area under the normal curve with a mean of 15,000 and a standard deviation of 4,500 lies to the left of the value of 16,170 generated by the Excel formula =NORM.INV(0.6026, 15000, 4500). Thus, the RAND function generates a percentage of the area under the normal curve, and then the NORM.INV function generates the corresponding value such that this randomly generated percentage lies to the left of this value.

Now suppose that the random number produced by the RAND function is 0.3551. Applying equation (13.6) then results in Demand =NORM.INV(0.3551, 15000, 4500) = 13,328 units. This matches intuition because half of this normal distribution lies below the mean of 15,000 and half lies above it, and so values generated by RAND() less than 0.5 result in values of first-year demand below the average of 15,000 units, and values generated by RAND() above 0.5 correspond to values of first-year demand above the average of 15,000 units.

Now that we know how to randomly generate values for the random variables (direct labor cost, parts cost, first-year demand) from their respective probability distributions, we modify the spreadsheet by adding this information. The static values in Figure 13.1 for these parameters in cells B6, B7, and B8 are replaced with cell formulas that will randomly generate values whenever the spreadsheet is recalculated (as shown in Figure 13.7). Cell B6 uses a random number generated by the RAND function in cell C6 and looks up the corresponding direct labor cost per unit by applying the XLOOKUP function to the table in cells A16:C20 (which corresponds to Table 13.1). Cell B7 executes equation (13.4) using a random number generated by the RAND function in cell C7 and references to the lower bound and upper bound of the uniform distribution of the parts cost in cells B23 and B24, respectively.[2] Cell B8 executes equation (13.6) a random number generated by the RAND function in cell C8 and

The XLOOKUP function is discussed in more detail in Chapter 12. Older versions of Excel may not recognize the XLOOKUP function, in which case VLOOKUP should be used.

Figure 13.6 Generation of Value for First-Year Demand Corresponding to Random Number 0.6026



[2]Technically, random variables modeled with continuous probability distributions should be appropriately rounded to avoid modeling error. For example, the simulated values of parts cost per unit should be rounded to the nearest penny. To simplify exposition, we do not worry about the small amount of error that occurs in this case. To model these random variables more accurately, the formula in cell B7 should be =ROUND(B22+(B23−B22)*RAND(),2).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.7 Sanotronics Formula Worksheet Demonstrating the Generation of Random Variables

	A	B	C	D
1	**Sanotronics**			
2				
3	**Parameters**			
4	Selling Price per Unit	249		
5	Administrative & Advertising Cost	1000000	Random Number	
6	Direct Labor Cost Per Unit	=XLOOKUP(C6,C16:C20,A16:A20,"error",1,2)	=RAND()	
7	Parts Cost Per Unit	=B23+(B24-B23)*C7	=RAND()	
8	Demand	=NORM.INV(C8,D23,D24)	=RAND()	
9				
10	**Model**			
11	Profit	=((B4-B6-B7)*B8)-B5		
12	Profit ($ million)	=B11/1000000		
13				
14	Direct Labor Cost			
15	Cost per Unit	Probability	Cumulative Probability	
16	43	0.1	=SUM($B$16:B16)	
17	44	0.1	=SUM($B$16:B17)	
18	45	0.4	=SUM($B$16:B18)	
19	46	0.2	=SUM($B$16:B19)	
20	47	0.2	=SUM($B$16:B20)	
21				
22	Parts Cost (Uniform)		Demand (Normal)	
23	Lower Bound	80	Mean	15000
24	Upper Bound	100	Standard Deviation	4500

MODEL *file*

sanotronicsblank

These steps iteratively select the simulation trial number from the range A27 through A1026 and substitute it into the blank cell selected in step 4 (D1). This substitution has no bearing on the spreadsheet, but it forces Excel to recalculate the spreadsheet each time, thereby generating new random numbers with the RAND functions in cells C6, C7, and C8.

using references to the mean and standard deviation of the normal distribution of the first-year demand in cells D23 and D24, respectively.[3]

Executing Simulation Trials with Excel

Each trial in the simulation involves randomly generating values for the random variables (direct labor cost, parts cost, and first-year demand) and computing profit. To facilitate the execution of multiple simulation trials, we use Excel's Data Table functionality in an unorthodox, but effective, manner. To set up the spreadsheet for the execution of 1,000 simulation trials, we construct a skeleton of a table as shown in cells A26 through E1026 in Figure 13.8. The cell range A27:A1026 numbers the 1,000 simulation trials (only the first 10 rows are shown in Figure 13.9). Cells B27:E27 contain references to the cells containing the simulated values of Direct Labor Cost, Parts Cost per Unit, Demand, and Profit ($ million) in cells B6, B7, B8, and B12, respectively. To populate the table of simulation trials in cells A27 through E1026, we execute the following steps (illustrated by Figure 13.9):

Step 1. Select cell range A27:E1026
Step 2. Click the **Data** tab in the Ribbon
Step 3. Click **What-If Analysis** in the **Forecast** group and select **Data Table…**

[3]In addition to being a continuous distribution that technically requires rounding when applied to discrete phenomena (like units of medical device demand), the normal distribution also allows negative values. The probability of a negative value is quite small in the case of first-year demand, and we simply ignore the small amount of modeling error for the sake of simplicity. To model first-year demand more accurately, the formula in cell B8 should be =MAX(ROUND(NORM.INV(RAND(),D22, D23),0),0).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.8 Setting Up Data Table for Sanotronics Simulation

	A	B	C	D	E
1	**Sanotronics**				
2					
3	**Parameters**				
4	Selling Price per Unit	249			
5	Administrative & Advertising Cost	1000000	Random Numbers		
6	Direct Labor Cost Per Unit	=XLOOKUP(C6,C16:C20,A16:A20,"error",1,2)	=RAND()		
7	Parts Cost Per Unit	=B23+(B24-B23)*C7	=RAND()		
8	Demand	=NORM.INV(C8,D23,D24)	=RAND()		
9					
10	**Model**				
11	Profit	=((B4-B6-B7)*B8)-B5			
12	Profit ($ million)	=B11/1000000			
13					
14	Direct Labor Cost				
15	Cost per Unit	Probability	Cumulative Probability		
16	43	0.1	=SUM($B$16:B16)		
17	44	0.1	=SUM($B$16:B17)		
18	45	0.4	=SUM($B$16:B18)		
19	46	0.2	=SUM($B$16:B19)		
20	47	0.2	=SUM($B$16:B20)		
21					
22	Parts Cost (Uniform)		Demand (Normal)		
23	Lower Bound	80	Mean	15000	
24	Upper Bound	100	Standard Deviation	4500	
25					
26	Simulation Trial	Direct Labor Cost Per Unit	Parts Cost Per Unit	Demand	Profit ($ million)
27	1	=B6	=B7	=B8	=B12
28	2				
29	3				
30	4				
31	5				
1026	1000				

Step 4. When the **Data Table** dialog box appears, leave the **Row input cell:** box blank and enter any empty cell in the spreadsheet (e.g., *D1*) into the **Column input cell:** box

Step 5. Click **OK**

Figure 13.10 shows the results of a set of 1,000 simulation trials. After executing the simulation with the Data Table, each row in this table corresponds to a distinct simulation trial consisting of different values of the random variables. In Trial 1 (row 27 in the spreadsheet), we see that the direct labor cost is $45 per unit, the parts cost is $81.53 per unit, and first-year demand is 22,003 units, resulting in profit of $1.695 million. In Trial 2 (row 28 in the spreadsheet), we observe random variables of $45 for the direct labor cost, $99.42 for the parts cost, and 23,282 for first-year demand. These values result in a simulated profit of $1.435 million on the second simulation trial. We note that every time the spreadsheet recalculates (by pressing the F9 key), new random values are generated by the RAND functions resulting in a new set of simulation trials.

Measuring and Analyzing Simulation Output

The analysis of the output observed over a set of simulation trials is a critical part of a simulation process. For the collection of simulation trials, it is helpful to compute descriptive statistics such as sample count, minimum sample value, maximum sample value, sample mean,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.9 Executing 1,000 Simulation Trials of Sanotronics Model

	A	B	C	D	E
1	**Sanotronics**				
2					
3	**Parameters**				
4	Selling Price per Unit	$249.00			
5	Administrative & Advertising Cost	$1,000,000	Random Number		
6	Direct Labor Cost Per Unit	$46.00	0.7328		
7	Parts Cost Per Unit	$82.06	0.1028		
8	Demand	9,728	0.1207		
9					
10	**Model**				
11	Profit	$176,507			
12	Profit ($ million)	$0.177			
13					
14	Direct Labor Cost				
15	Cost per Unit	Probability	Cumulative Probability		
16	$43	0.1	0.1		
17	$44	0.1	0.2		
18	$45	0.4	0.6		
19	$46	0.2	0.8		
20	$47	0.2	1		
21					
22	Parts Cost (Uniform)		Demand (Normal)		
23	Lower Bound	$80.00	Mean	15,000	
24	Upper Bound	$100.00	Standard Deviation	4,500	
25					
26	Simulation Trial	Direct Labor Cost Per Unit	Parts Cost Per Unit	Demand	Profit ($ million)
27	1	$46.00	$82.06	9,728	$0.177
28	2				
29	3				
30	4				
31	5				
32	6				
33	7				
34	8				
35	9				
36	10				

Data Table ? ×
Row input cell:
Column input cell: $D$1
OK Cancel

and sample standard deviation. To compute these statistics for the Sanotronics example, we use the following Excel functions:

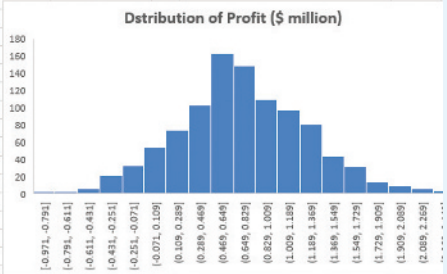
MODEL file
sanotronics

Cell H27	=*COUNT(E27:E1026)*
Cell H28	=*MIN(E27:E1026)*
Cell H29	=*MAX(E27:E1026)*
Cell H30	=*AVERAGE(E27:E1026)*
Cell H31	=*STDEV.S(E27:E1026)*

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.10 Output from Sanotronics Simulation

	A	B	C	D	E	F	G	H
1	**Sanotronics**							
2								
3	**Parameters**							
4	Selling Price per Unit	$249.00						
5	Administrative & Advertising Cost	$1,000,000	Random Number					
6	Direct Labor Cost Per Unit	$45.00	0.5003					
7	Parts Cost Per Unit	$81.53	0.0764					
8	Demand	22,003	0.9402					
9								
10	**Model**							
11	Profit	$1,694,768						
12	Profit ($ million)	$1.695						
13								
14	Direct Labor Cost							
15	Cost per Unit	Probability	Cumulative Probability					
16	$43	0.1	0.1					
17	$44	0.1	0.2					
18	$45	0.4	0.6					
19	$46	0.2	0.8					
20	$47	0.2	1					
21								
22	Parts Cost (Uniform)		Demand (Normal)					
23	Lower Bound	$80.00	Mean	15,000				
24	Upper Bound	$100.00	Standard Deviation	4,500				
25								
26	Simulation Trial	Direct Labor Cost Per Unit	Parts Cost Per Unit	Demand	Profit ($ million)		Profit ($ million) Summary Statistics	
27	1	$45.00	$81.53	22,003	$1.695		Count	1000
28	2	$45.00	$99.42	23,282	$1.435		Minimum	-$0.971
29	3	$47.00	$82.96	18,137	$1.159		Maximum	$2.317
30	4	$45.00	$85.59	18,608	$1.203		Mean	$0.726
31	5	$44.00	$81.04	19,313	$1.394		Standard Deviation	$0.526
32	6	$43.00	$93.28	20,401	$1.300			
33	7	$46.00	$99.54	21,333	$1.207		95% C.I. on Mean Profit	$0.693
34	8	$43.00	$90.81	13,031	$0.501			$0.758
35	9	$45.00	$83.16	12,672	$0.531			
36	10	$47.00	$96.96	22,364	$1.349		95% P.I. on Profit	-$0.300
37	11	$45.00	$86.63	13,542	$0.590			$1.810
38	12	$46.00	$90.95	18,557	$1.079			
39	13	$45.00	$93.27	14,076	$0.559		P(Profit < $0)	0.089
40	14	$45.00	$90.03	18,080	$1.061		Standard Error of Proportion	0.009
41	15	$45.00	$84.27	11,617	$0.391			
42	16	$45.00	$98.90	15,223	$0.600		95% C.I. on P(Profit < $0)	0.071
43	17	$44.00	$84.45	13,422	$0.618			0.107



As shown in Figure 13.10, for this sample of 1,000 observations, profit ranges from −$0.971 million to $2.317 million. The sample mean profit is $0.726 million and the sample standard deviation is $0.526 million.

To visualize the distribution of profit on which these descriptive statistics are based, we create a histogram with the following steps:

Step 1. Select cells E27:E1026
Step 2. Click the **Insert** tab on the Ribbon
Step 3. Click the **Insert Statistic Chart button** in the **Charts** group

When the list of statistic charts appears, select **Histogram**

Figure 13.10 shows that the distribution of profit values appears unimodal and fairly symmetric, with a large proportion of values between $0.1 million and $1.3 million. Only a few trials out of 1,000 resulted in a loss of more than $0.6 million or a profit greater than $1.9 million.

In comparing the simulation approach to the what-if approach, we observe that much more information is obtained using simulation. Recall from the what-if analysis in Section 13.1, we learned that the base-case scenario projected a profit of $0.71 million. The worst-case scenario projected a loss of $1 million, and the best-case scenario projected a profit of $2.591 million. From the 1,000 trials of the simulation that have been run, we see that extremes such as the worst- and best-case scenarios, although possible,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

are unlikely. Indeed, the advantage of simulation for risk analysis is the information it provides on the likelihood of output values. For the assumed distributions of the direct labor cost, parts cost, and demand, we can estimate the probability of a loss, how the profit values are distributed over their range, and what profit values are most likely.

When pressing the **F9** key to generate a new set of 1,000 simulation trials, we observe that the distribution of the 1,000 profit values and the summary statistics vary. In particular, the sample mean profit fluctuates for each new set of simulation trials. To account for this sampling error, we can construct a confidence interval on the mean profit. Recall that the general formula for a confidence interval is: point estimate +/− margin of error. To compute a 95% confidence interval on mean profit for the Sanotronics example, we use the following Excel functions:

Confidence intervals are discussed in more detail in Chapter 7.

Cell H33 `=H30 − CONFIDENCE.T(0.05, H31, H27)`
Cell H34 `=H30 + CONFIDENCE.T(0.05, H31, H27)`

Cells H33 and H34 compute the lower and upper limits of a 95% confidence interval of the mean profit. To compute the margin of error for this interval estimate, the Excel CONFIDENCE.T function requires three arguments: the significance level (1 − confidence level), the sample standard deviation, and the sample size.

Figure 13.10 shows a 95% confidence interval on the mean profit ranging from $0.693 million to $0.758 million. The width of this confidence interval reflects the degree of sampling error in the sample mean that results from using only 1,000 trial values.

A common misinterpretation is to relate the 95% confidence interval on the mean profit to the profit distribution of the 1,000 simulated profit values displayed in Figure 13.10. Looking at the profit distribution it is clear that 95% of the values do not lie between $0.693 million and $0.758 million (the interval corresponding to the 95% confidence interval on the mean profit). The 95% confidence interval relates only to the confidence we have in the estimation of the mean profit, not the likelihood of an individual profit observation. If we desire an interval that contains 95% of the profit observations, we can construct this by using the Excel PERCENTILE.EXC function. To compute a 95% probability interval on profit for the Sanotronics example, we use the following Excel functions:

Cell H36 `=PERCENTILE.EXC(E27:E1026,0.025)`
Cell H37 `=PERCENTILE.EXC(E27:E1026,0.975)`

For this set of 1,000 profit values, −$0.300 and $1.810 are the lower and upper limits of an interval estimating the range that is 95% likely to contain the profit outcome for Sanotronics.

In addition to estimating the mean profit, simulation can be used to estimate the likelihood of an event of interest. In particular, obtaining an estimate of the probability of a specified negative event is an important part of risk analysis. For the Sanotronics example, we define the occurrence of Sanotronics losing money in its first year as the event of interest. To estimate the probability of Sanotronics' profit being less than zero, we use the following Excel functions to compute the sample proportion and the sample standard error of the proportion:

The probability of any specified event can be estimated by appropriately changing the value of the second argument in the COUNTIF function in cell H39 or by using a COUNTIFS function if multiple conditions are necessary.

Cell H39 `=COUNTIF(E27:E1026,"<0")/H27`
Cell H40 `=SQRT(H39*(1-H39)/H27)`

Cell H39 computes the ratio of the number of trials whose profit is less than zero to the total number of trials. Cell H40 computes the sample standard error of the proportion using the formula $\sqrt{(\overline{p}(1-\overline{p})/n)}$, where $\overline{p}$ is the sample proportion of observations satisfying a criterion (profit less than $0 in this case) and n is the sample size (1,000 in this case). The sample standard error of the proportion provides a measure of how much the sample proportion varies across different samples of 1,000 simulation trials.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

As shown in Figure 13.10, there is a sample proportion of 0.089 of the observations with negative profit and the sample standard error of this estimate is 0.009. Relying on the general formula for a confidence interval of point estimate +/− margin of error, we use the following Excel functions to compute a 95% confidence interval on the proportion:

Cell H42 $=H39 - (NORM.S.INV(0.975)*H40)$

Cell H43 $=H39 + (NORM.S.INV(0.975)*H40)$

Recall that NORM.S.INV(0.975) computes the value such that 2.5% of the area under the standard normal distribution lies in the right tail beyond this value.

Cells H42 and H43 compute the lower and upper limits of a 95% confidence interval on the proportion of observations with a negative profit. To compute the margin of error for this interval estimate, the sample standard error of the proportion (in cell H40) is multiplied by the z value corresponding to a 95% confidence level (as calculated by NORM.S.INV(0.975)). Figure 13.10 shows that a 95% confidence interval on the probability of a negative profit ranges from 0.071 to 0.107.

The simulation results help Sanotronics's management better understand the profit/loss potential of the new medical device. An estimated 0.071 to 0.107 probability of a loss with an estimated mean profit between $0.693 million and $0.758 million may be acceptable to management. On the other hand, Sanotronics might want to conduct further market research before deciding whether to introduce the product. In any case, the simulation results should be helpful in reaching an appropriate decision.

Notes + Comments

1. For every example, problem, and case in this chapter, we provide a MODELfile that contains the skeleton of the table of simulation trials and formulas for computing the summary statistics. This allows the user to focus on the construction of the logic for the simulation model and the modeling of the uncertain input quantities. The user then only needs to correctly reference the cell(s) containing the output measure(s) of interest and execute the 1,000 simulation trials via the Data Table tool in order to generate output to analyze.
2. In the preceding section, we showed how to generate values for random variables from a generic discrete distribution, a uniform distribution, and a normal distribution. Generating values for a normally distributed random variable required the use of the NORM.INV and RAND functions. When using the Excel formula =NORM.INV(RAND( ), m, s), the RAND( ) function generates a random number r between 0 and 1 and then the NORM.INV function identifies the smallest value k such that $P(X \leq k) \geq r$, where X is a normal random variable with mean m and standard deviation s. Similarly, the RAND function can be used with the Excel functions BETA.INV, BINOM.INV, GAMMA.INV, and LOGNORM.INV to generate values for a random variable with a beta distribution, binomial distribution, gamma distribution, and lognormal distribution, respectively. Using a different probability distribution for a random variable simply changes the relative likelihood of the random variable realizing certain values. The choice of probability distribution to use for a random variable should be based on historical data and knowledge of the analyst. In Appendix 13.1, we discuss several probability distributions and how to generate them with native Excel functions.
3. We can reduce the width of the confidence intervals associated with the sample mean and the sample proportion computed from a set of simulation trials by increasing the number of trials beyond 1,000. However, increasing the number of trials can begin to tax the computational capabilities of Excel. When more than 1,000 trials are necessary to reduce the sampling error, the analyst may want to restrict Excel to only update values upon a specific command rather than updating anytime the Enter key is pressed in Excel. This can be accomplished by choosing **File** from the Ribbon, clicking **Options**, choosing **Formulas**, and then changing the **Calculation options** to **Manual**. When this change is made, Excel will update values only when the **F9** key is pressed.

13.2 Inventory Policy Analysis for Promus Corp

In this section, we demonstrate how simulation can be used to evaluate an inventory policy for a product that has an uncertain demand. In our example, we consider Promus Corp, which sells wireless routers. Each router costs Promus $75 and the company sells it for $125.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Thus, Promus realizes a gross profit of \$125 − \$75 = \$50 for each router sold. Monthly demand for the router is uncertain, but Promus Corp has collected data to help characterize the uncertainty.

Promus receives monthly deliveries from its supplier and replenishes its inventory to a predetermined level at the beginning of each month. This predetermined inventory level is referred to as the replenishment level. If monthly demand is less than the replenishment level, then Promus must hold the unsold routers in inventory. Each unsold router held by Promus results in an inventory holding cost of \$15 because Promus must pay for the storage, insurance, and cost-of-capital for the inventory. However, if monthly demand is greater than the replenishment level, then a stock-out occurs which results in a shortage cost of \$30 being charged for each unit of demand that cannot be satisfied. Management would like to use a simulation model to determine the average monthly net profit resulting from using different replenishment levels.

MODEL file
promusblank

Spreadsheet Model for Promus

To evaluate different replenishment levels, we develop a simulation model that appropriately accounts for the relationships between the input parameters to compute the output measure(s) of interest. For Promus, the input parameters are the gross profit per unit (known to be \$50), the unit holding cost (\$15), the unit shortage cost (\$30), the monthly demand (uncertain, D), and the replenishment level (a decision, Q). The output measure of interest is the monthly net profit.

Having identified the input parameters and output measure, the next step is to establish computational logic that determines the output measure (monthly net profit) for given values of the input parameters. For Promus, the computation of monthly profit depends on the relative magnitude of the monthly demand and the replenishment level.

For a specified replenishment level (Q) and observed monthly demand (D), monthly net profit is calculated as follows:

$$\begin{aligned}\text{Monthly net profit} &= \text{Gross profit} - \text{holding cost} - \text{shortage cost}\\ &= 50 \times \min\{D, Q\} - 15 \times \max\{Q - D, 0\} - 30 \times \max\{D - Q, 0\}\end{aligned} \quad \textbf{(13.7)}$$

It is informative to analyze equation (13.7) for the two cases: $D \le Q$ and $D > Q$. When demand is less than or equal to the replenishment level ($D \le Q$), D units are sold, and an inventory holding cost of \$15 is incurred for each of the $Q - D$ units that remain in storage, and no shortage occurs. This results in:

$$\text{Monthly net profit} = 50 \times D - 15 \times (Q - D)$$

When demand is greater than the replenishment level ($D > Q$), Q units are sold, no inventory remains, and a shortage of $D - Q$ occurs. This results in:

$$\text{Monthly net profit} = 50 \times Q - 30 \times (D - Q)$$

For a replenishment level of 90 and monthly demand of 100, Figure 13.11 displays the Excel implementation of equation (13.7). The gross profit per unit, holding cost per unit, and shortage cost per unit data are entered into cells B4, B5, and B6. An observed demand of 100 units is entered into cell B7. The replenishment level (a controllable input) of 90 is entered into cell B10. The calculation of monthly net profit according to equation (13.7) is broken down into several separate calculations in cells B11 through B15.

Cell B11	*=MIN(B7, B10)*
Cell B12	*=B11*B4*
Cell B13	*=IF(B7<=B10, (B10−B7)*B5, 0)*
Cell B14	*=IF(B7>B10, (B7−B10)*B6, 0)*
Cell B15	*=B12−B13−B14*

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.11 Excel Worksheet for Promus Corp

	A	B
1	**Promus Corp**	
2		
3	**Parameters**	
4	Gross Profit per Unit	50
5	Holding Cost per Unit	15
6	Shortage Cost per Unit	30
7	Demand	100
8		
9	**Model**	
10	Replenishment Level (Q)	90
11	Sales	=MIN($B$7,B10)
12	Gross Profit	=B$11*$B$4
13	Holding Cost	=IF($B$7<=B10,(B10-$B$7)*$B$5,0)
14	Shortage Cost	=IF($B$7>B10,($B$7-B10)*$B$6,0)
15	Net Profit	=B12-B13-B14

	A	B
1	**Promus Corp**	
2		
3	**Parameters**	
4	Gross Profit per Unit	$50.00
5	Holding Cost per Unit	$15.00
6	Shortage Cost per Unit	$30.00
7	Demand	100
8		
9	**Model**	
10	Replenishment Level (Q)	90
11	Sales	90
12	Gross Profit	$4,500
13	Holding Cost	$0
14	Shortage Cost	$1,065
15	Net Profit	$3,435

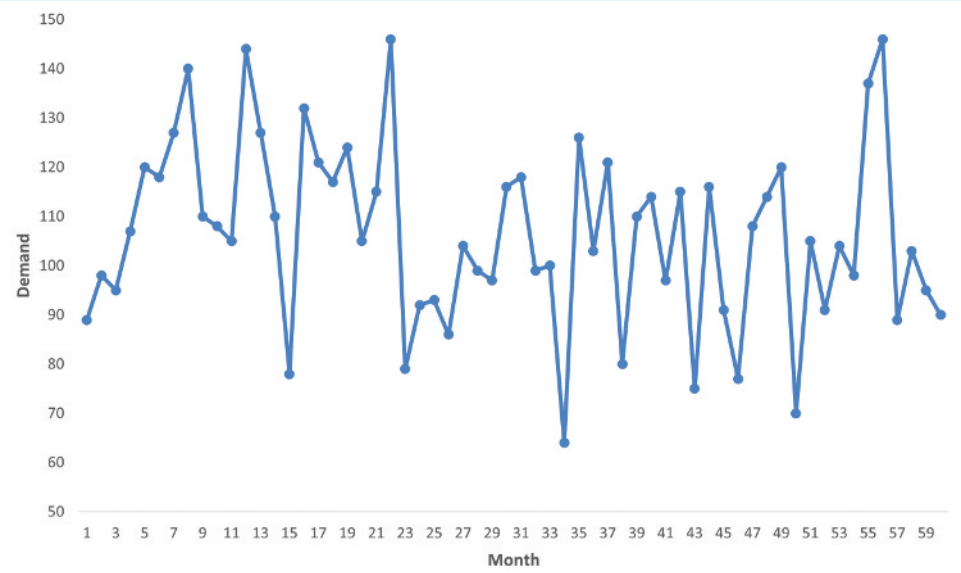
Generating Values for Promus Corp's Demand

With a spreadsheet model that correctly computes monthly net profit for given values of the inputs (including a specified replenishment level and an observed demand), the next step is to characterize the demand uncertainty and randomly generate values for demand in a manner that reflects the relative likelihood of future monthly demand values.

To gain a better understanding of its router business, Promus has recorded the demand for its routers for the past 60 months. Figure 13.12 plots the monthly demand over time and illustrates that, while router demand varies from month to month, there is no detectable

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

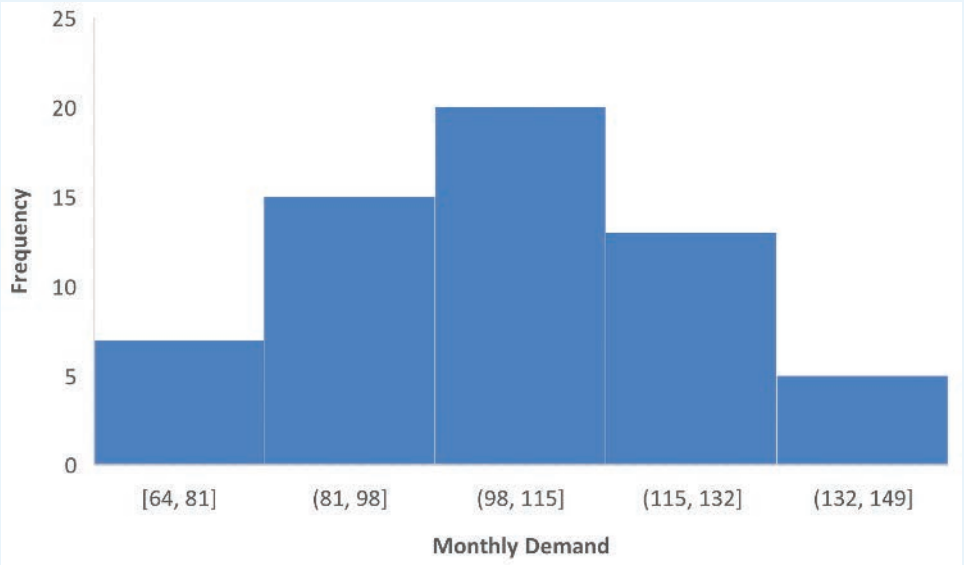
Figure 13.12 Router Demand Over Time



pattern in the variation. That is, there does not appear to be any trend or seasonality in the demand. Therefore, Promus feels comfortable in treating router demands from month to month as independent quantities.

Figure 13.13 characterizes the relative likelihood of monthly demand values with a histogram of the 60 monthly demand observations. As Figure 13.13 illustrates, the distribution of monthly demand is somewhat bell-shaped, suggesting that the normal distribution could

Figure 13.13 Distribution of Monthly Router Demand



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

be a good fit to these data. Due to the wide range of possible values of monthly demand, using a continuous distribution such as the normal distribution is appropriate.[4] To describe the normal distribution fitting these data, we compute the sample mean ($\bar{x} = 106$) and sample standard deviation ($s = 19$) of these 60 observations.

We can randomly generate a value for monthly demand described by a normal distribution with a mean of 106 units and a standard deviation of 19 units using the formula =NORM.INV(RAND(), 106, 19). In Figure 13.14, we modify the spreadsheet model by replacing the static value of demand in cell B7 with the formula *=NORM.INV(C7, B20, B21)*, where cell C7 contains a random number generated by RAND, and cells B20 and B21 contain the values for the sample mean and sample standard deviation, respectively.

Promus would like to compare the average monthly profit corresponding to a replenishment level of 90 units to the average monthly profit corresponding to a replenishment level of 110 units. To facilitate this comparison, in cells C11:C15 displayed in Figure 13.14, we implement the same logic for a replenishment level of 110 that we did for a replenishment level of 90 in cells B11:B15. We note that the computations for these two replenishment levels all use the same input data (cells B4 through B7)[5] and only differ by the replenishment level applied (cells B10 and C10). In cell B17, we compute the difference between the net profit when $Q = 110$ and the net profit when $Q = 90$ for the same value of observed demand in cell B7.

Figure 13.14 Modeling Demand as a Normal Random Variable and Computing Profit Difference Between $Q = 90$ and $Q = 110$

	A	B	C
1	**Promus Corp**		
2			
3	**Parameters**		
4	Gross Profit per Unit	50	
5	Holding Cost per Unit	15	
6	Shortage Cost per Unit	30	Random Number
7	Demand	=NORM.INV(C7,B20,B21)	=RAND()
8			
9	**Model**		
10	Replenishment Level (Q)	90	110
11	Sales	=MIN($B$7,B10)	=MIN($B$7,C10)
12	Gross Profit	=B$11*$B$4	=C$11*$B$4
13	Holding Cost	=IF($B$7<=B10,(B10-$B$7)*$B$5,0)	=IF($B$7<=C10,(C10-$B$7)*$B$5,0)
14	Shortage Cost	=IF($B$7>B10,($B$7-B10)*$B$6,0)	=IF($B$7>C10,($B$7-C10)*$B$6,0)
15	Net Profit	=B12-B13-B14	=C12-C13-C14
16			
17	$Profit_{Q=110}$ - $Profit_{Q=90}$	=C15-B15	
18			
19	**Demand (Normal)**		
20	Mean	106	
21	Standard Deviation	19	

[4]Note that technically demand values for routers can only take integer values. The formula in Excel can be modified to require this using the Excel ROUND function, but for distributions with a relatively large mean we generally use a continuous random variable to model demand.

[5]The use of the same set of random values for demand to evaluate both replenishment levels is an example of a general technique known as common random numbers. This is not strictly required for problems such as this, but it reduces the variance in the output measures to allow for a more stable comparison of different replenishment levels. For more information on common random numbers, see Law, A.M., *Simulation Modeling and Analysis*, 5th edition, McGraw-Hill, 2014.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Executing Simulation Trials and Analyzing Output

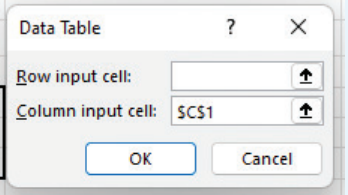
Each trial in the simulation involves randomly generating a value for the corresponding month's demand, and then computing the difference in the net profit when using a replenishment level of 110 versus a replenishment level of 90. To execute a simulation experiment of 1,000 trials and analyze the output, we structure the spreadsheet in a similar manner as the Sanotronics example in Section 13.1. Figure 13.15 shows a portion of the table of simulation trials. The cell range from A24 through C1023 has been prepared to hold the set of 1,000 simulation trials. Cell range A24:A1023 numbers the rows that will correspond to the 1,000 simulation trials (only the first five rows are shown in Figure 13.15). Cells B24 and C24 contain Excel formulas referencing the random variable (monthly demand) and the output measure (difference in the net profit resulting from $Q = 110$ and $Q = 90$).

To populate the table of simulation trials in the Model worksheet, we execute the following steps:

Step 1. Select cell range A24:C1023
Step 2. Click the **Data** tab in the Ribbon
Step 3. Click **What-If Analysis** in the **Forecast** group and select **Data Table…**
Step 4. When the **Data Table** dialog box appears, leave the **Row input cell:** box blank and enter any empty cell in the spreadsheet (e.g., *C1*) into the **Column input cell:** box
Step 5. Click **OK**

Figure 13.15 Executing 1,000 Simulation Trials of Promus Model

	A	B	C
1	**Promus Corp**		
2			
3	**Parameters**		
4	Gross Profit per Unit	50	
5	Holding Cost per Unit	15	
6	Shortage Cost per Unit	30	Random Number
7	Demand	=NORM.INV(RAND(),B20,B21)	=RAND()
8			
9	**Model**		
10	Replenishment Level (Q)	90	110
11	Sales	=MIN($B$7,B10)	=MIN($B$7,C10)
12	Gross Profit	=B$11*$B$4	=C$11*$B$4
13	Holding Cost	=IF($B$7<=B10,(B10-$B$7)*$B$5,0)	=IF($B$7<=C10,(C10-$B$7)*$B$5,0)
14	Shortage Cost	=IF($B$7>B10,($B$7-B10)*$B$6,0)	=IF($B$7>C10,($B$7-C10)*$B$6,0)
15	Net Profit	=B12-B13-B14	=C12-C13-C14
16			
17	$Profit_{Q=110} - Profit_{Q=90}$	=C15-B15	
18			
19	**Demand (Normal)**		
20	Mean	106	
21	Standard Deviation	19	
22			
23	Simulation Trial	Demand	$Profit_{Q=110} - Profit_{Q=90}$
24	1	=B7	=B17
25	2		
26	3		
27	4		
28	5		



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.16 shows the results of a set of 1,000 simulation trials. After executing the simulation with the Data Table, each row of this table corresponds to a distinct simulation trial consisting of different values of monthly demand. We see in the simulated month corresponding to Trial 1 that demand is 116 units and a replenishment level of 110 earns $1,600 more than a replenishment level of 90. In Trial 2, monthly demand is 102 and a replenishment level of 110 earns $880 more than a replenishment level of 90.

Similar to the Sanotronics example in Section 13.1, we compute sample statistics, 95% confidence intervals on the mean and the proportion, and a 95% probability interval on individual outcomes based on the 1,000 simulation trials. The formulas of these summary statistics in the cell range F24:F40 are analogous to the Sanotronics example with the exception of the calculation of the sample proportion in cell F36. In cell F36, the probability that the profit from $Q = 110$ will exceed the profit from $Q = 90$ is computed as:

Cell F36 $=COUNTIF(C24:C1023, ">0")/F24$

Figure 13.16 shows that based on this set of 1,000 simulation trials, there is a mean difference of $858 between the net profit generated by a replenishment level of 110 versus a replenishment level of 90. Because the 95% confidence interval on the mean difference, [$809, $906], does not contain zero, we have reason to believe that this difference is statistically significant. Furthermore, there is an estimated probability of 0.746 that a replenishment level of 110 will generate a larger monthly profit than a replenishment level of 90. The 95% confidence interval on this proportion is [0.719, 0.773] which further supports the notion that $Q = 110$ is statistically superior to $Q = 90$.

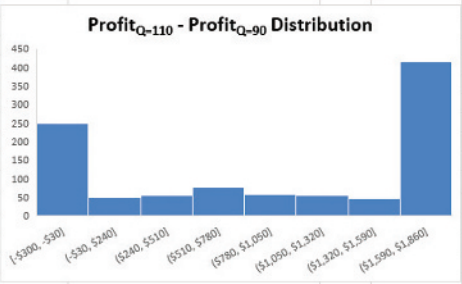
*To obtain a static set of values over the 1,000 simulation trials, replace the dynamic table by selecting the range A24:C1023, **Copy** and then **Paste Values** in the same range.*

We note that a different set of 1,000 simulation trials can be generated by pressing the **F9** key, and this will result in different values of the summary statistics because these will now

MODEL file
promus

Figure 13.16 Output Worksheet for Promus Corp

	A	B	C	D	E	F	G
1	**Promus Corp**						
2							
3	**Parameters**						
4	Gross Profit per Unit	$50.00					
5	Holding Cost per Unit	$15.00					
6	Shortage Cost per Unit	$30.00	Random Number				
7	Demand	116	0.4777				
8							
9	**Model**						
10	Replenishment Level (Q)	90	110				
11	Sales	90	110				
12	Gross Profit	$4,500	$5,500				
13	Holding Cost	$0	$0				
14	Shortage Cost	$767	$167				
15	Net Profit	$3,733	$5,333				
16							
17	$Profit_{Q=110} - Profit_{Q=90}$	$1,600					
18							
19	**Demand (Normal)**						
20	Mean	106					
21	Standard Deviation	19					
22							
23	Simulation Trial	Demand	$Profit_{Q=110} - Profit_{Q=90}$		$Profit_{Q=110} - Profit_{Q=90}$ Summary Statistics		
24	1	116	$1,600		Count	1000	
25	2	102	$880		Minimum	-$300	
26	3	112	$1,600		Maximum	$1,600	
27	4	128	$1,600		Mean	$858	
28	5	129	$1,600		Standard Deviation	$784	
29	6	109	$1,495				
30	7	84	-$300		95% C.I. on Mean Difference	$809	
31	8	95	$180			$906	
32	9	113	$1,600				
33	10	91	-$232		95% P.I. on Difference	-$300	
34	11	144	$1,600			$1,600	
35	12	148	$1,600				
36	13	87	-$300		P(Difference > $0)	0.746	
37	14	103	$981		Standard Error of Proportion	0.014	
38	15	104	$1,047				
39	16	73	-$300		95% C.I. on P(Difference > $0)	0.719	
40	17	104	$986			0.773	



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

be based on a different sample. By pressing the **F9** key and observing how much the output statistics vary, the analyst can gauge how much sampling error exists in the output statistics. Furthermore, the 95% confidence interval on the mean difference and the 95% confidence interval on the $P(\text{Difference} > 0)$ reflect the degree of the sampling error. Wider confidence intervals reflect more uncertainty in the accuracy of the sample mean and sample proportion. If we would press the **F9** key 100 times to create 100 different samples of 1,000 trials, we would expect 95 of the corresponding 100 confidence intervals on the mean difference to contain the true mean difference in the net profit for these two replenishment levels. Similarly, we would expect 95 of the 100 confidence intervals on the proportion of months that $Q = 110$ earns more profit than $Q = 90$ to contain the true probability that $Q = 110$ earns more profit than $Q = 90$ in a month.

Different pairs of replenishment levels can be compared by changing the replenishment levels in cells B10 and C10. When comparing replenishment levels that are very similar, e.g., $Q = 100$ and $Q = 105$, the sampling error in the simulation experiment may be too large to establish a statistically significant difference. In general, increasing the number of trials in a simulation experiment will decrease the variability in the summary statistics from one sample of simulation trials to the next. Therefore, if we wish to decrease the sampling error in the output statistics, we should increase the number of simulation trials and re-execute the simulation experiment. However, at this point, we must also discern if the difference in net profit is practically significant even if we can establish its statistical significance by increasing the number of simulation trials. Practical significance must be determined by the decision maker and consider such factors as how expensive it is to implement a new inventory policy.

13.3 Simulation Modeling for Land Shark Inc.

Land Shark Inc., a real estate company, purchases properties that it develops and then resells. In the past, Land Shark has successfully acquired properties via first-price sealed-bid auctions involving commercial and residential properties. In such auctions, each bidder submits a single concealed bid. The submitted bids are then compared, and the party with the highest bid wins the property and pays the bid amount. In case of a tie (a rare occurrence), a coin flip decides the winner.

MODEL file
landsharkblank

Land Shark has been reviewing upcoming property auctions and has identified a commercial property of interest. Land Shark estimates the value of this property to be $1,389,000. Using bidding data disclosed to the public, Land Shark has maintained a file summarizing 56 previous auctions that it believes are similar to the upcoming property auction. Table 13.2

Table 13.2 Portion of Bid Data on Commercial Property Auctions

	Bid Amount (as a Fraction of Estimated Property Value)							
Property No.	Bid 1	Bid 2	Bid 3	Bid 4	Bid 5	Bid 6	Bid 7	Bid 8
1	0.830	0.797	0.833	0.878	0.839	0.843		
2	0.835	0.823	0.781	0.892	0.767	0.787		
3	0.763	0.862	0.814	0.895				
4	0.771	0.859	0.867	0.850	0.833			
5	0.836	0.898	0.831	0.897	0.831	0.657	0.846	
6	0.850	0.863	0.825	0.910	0.848			
7	0.890	0.820	0.874	0.877	0.818			
8	0.804	0.881	0.786	0.884	0.773	0.819	0.824	
9	0.819	0.851	0.786	0.896	0.784	0.792		
10	0.860	0.756	0.876	0.887	0.866			
11	0.880	0.834	0.831	0.871	0.857	0.759		
12	0.810	0.870						
13	0.887	0.716	0.817	0.9	0.869	0.885	0.856	0.761

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

displays bid data for a portion of Land Shark's data. The data for all 56 auctions is in the *Auctions* worksheet of the file *landsharkblank*. Because the property value up for sale varies between auctions, Land Shark expresses the submitted bid amounts as fractions of the respective property's value to make the bids in different auctions comparable. These bid percentages can be converted into a bid amount (in dollars) by multiplying the bid percentage by the estimated value of the property under auction. Land Shark is considering a bid of $1,229,000 and would like to evaluate its chances of winning the upcoming auction with this bid.

Spreadsheet Model for Land Shark

To evaluate Land Shark's chances of winning the auction, we develop a simulation model. Our first step in modeling the upcoming property auction is to identify the input parameters and output measures. The next step is to develop a spreadsheet model that correctly computes the values of the output measures given static values of the input parameters. Then we prepare the spreadsheet model for simulation analysis by replacing the static values of the input parameters that Land Shark does not know with certainty with probability distributions of possible values.

The relevant input parameters for the upcoming auction are the estimated value of the property, the number of bidders competing against Land Shark, the bid amounts submitted by the competitors, and Land Shark's bid amount. Land Shark is certain about its estimate that the property is worth $1,389,000. Furthermore, Land Shark controls its bid amount and it would like to evaluate a bid amount of $1,229,000. However, Land Shark is uncertain about the number of competing bidders and the bid amounts submitted by these competitors.

The output measures in which we are interested are whether Land Shark wins the simulated auction given its specified amount and Land Shark's net return. If Land Shark wins the auction, its return is computed as the difference between the estimated value of the property and its bid amount. If Land Shark does not win the auction, its return is $0.

To understand how to construct the logic for determining whether Land Shark wins an auction and its return from the auction, let's first consider the data and calculations necessary for a scenario with static values for the uncertain input parameters (number of competing bidders and the submitted bid amounts). Based on Land Shark's data on the past 56 auctions, the number of competitor bids ranges from two to eight. Therefore, the spreadsheet model needs to accommodate as many as eight different bid amounts submitted by competitors. Suppose those eight competitor bid amounts (as a percentage of the property's estimated value) are 0.887, 0.716, 0.818, 0.900, 0.869, 0.885, 0.856, and 0.761. However, it is possible that not all eight of these bid amounts will be submitted for an auction. Suppose only four competitors decide to submit bids in the auction. Then we only want to consider four of the eight bid amounts. If the bid amounts are listed in a random order (which they are in this case), we can just select the first four bid amounts and ignore the last four. In this case, the four competing bid amounts (expressed in dollars) are: (0.887)($1,389,000) = $1,232,043; (0.716)($1,389,000) = $994,524; (0.818)($1,389,000) = $1,136,202; and (0.900)($1,389,000) = $1,250,100. The largest competing bid amount is then the maximum of these four bid amounts, which is $1,250,100. We compare Land Shark's bid ($1,229,000) to the largest competing bid ($1,250,100) and observe that in this scenario, Land Shark does not win the auction, so its return is $0.

In the example in the previous paragraph, we determined the largest bid from four competitors by considering only the first four competitor bids and ignoring the last four. In general, the number of competitor bids is uncertain and varies from two to eight. Therefore, we need to devise a spreadsheet model that will correctly compute the largest competing bid amount from among a varying number of bids. Figure 13.17 shows the formula view and value view of the spreadsheet implementing one way to model the problem. Cell B4 contains the estimated value of the property (Land Shark is certain of this value) and cell B5 contains a value for the number of bidders (Land Shark is uncertain of this value). Cell range B8:B15 contains the values of eight possible competing bids expressed as fractions of the property's estimated value (Land Shark is uncertain of these values). Cells C8 through C15 express the respective bid fractions in cells B8 through B15 as dollar amounts using the IF function to determine if the bid should be considered or effectively eliminated. If a bid index (from the range A8:A15)

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.17 Base Spreadsheet Model for Land Shark

	A	B	C
1	**Land Shark**		
2			
3	**Parameters**		
4	Estimated Value	1389000	
5	Number of Bidders	4	
6			
7	Bid Index	Bid Fraction	Bid Amount
8	1	0.887	=IF(A8>$B$5,0,B8*$B$4)
9	2	0.716	=IF(A9>$B$5,0,B9*$B$4)
10	3	0.818	=IF(A10>$B$5,0,B10*$B$4)
11	4	0.9	=IF(A11>$B$5,0,B11*$B$4)
12	5	0.869	=IF(A12>$B$5,0,B12*$B$4)
13	6	0.885	=IF(A13>$B$5,0,B13*$B$4)
14	7	0.856	=IF(A14>$B$5,0,B14*$B$4)
15	8	0.761	=IF(A15>$B$5,0,B15*$B$4)
16			
17	**Model**		
18	Land Shark Bid Amount	1229000	
19	Largest Competitor Bid	=MAX(C8:C15)	
20	Land Shark Win Auction?	=IF(B18>B19,1,0)	
21	Land Shark Return	=B20*(B4-B18)	

	A	B	C
1	**Land Shark**		
2			
3	**Parameters**		
4	Estimated Value	$1,389,000	
5	Number of Bidders	4	
6			
7	Bid Index	Bid Fraction	Bid Amount
8	1	0.887	$1,232,043
9	2	0.716	$994,524
10	3	0.818	$1,136,202
11	4	0.900	$1,250,100
12	5	0.869	$0
13	6	0.885	$0
14	7	0.856	$0
15	8	0.761	$0
16			
17	**Model**		
18	Land Shark Bid Amount	$1,229,000	
19	Largest Competitor Bid	$1,250,100	
20	Land Shark Win Auction?	0	
21	Land Shark Return	$0	

exceeds the realized number of bidders in cell B5, the corresponding bid amount in the cell range C8:C15 is set to $0, otherwise the bid amount is computed. For example, consider the formula in cell C8, =IF(A8>B5, 0, B8*B4). This formula compares the bid index in cell A8 to the number of bidders in cell B5, and if the bid index exceeds the number of bidders, a bid amount of $0 is calculated so that the bid is not considered. Otherwise, the bid amount is calculated by multiplying the bid fraction by the estimated value of the property.

The IF function is discussed in more detail in Chapter 12.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Cell B18 contains Land Shark's bid amount. Cell B19 computes the largest competing bid by taking the maximum value over the range C8:C15. Land Shark tracks two output measures: whether it wins the auction and the return from the auction. By comparing Land Shark's bid amount in cell B18 to the largest competitor bid in cell B19, the formula =IF(B18>B19,1,0) in cell B20 returns a value of 1 if Land Shark wins the auction and a value of 0 if Land Shark loses the auction. The value of 1 or 0 in Cell B20 to denote a Land Shark win or loss allows the simulation model to easily count the number of times Land Shark wins the auction over a set of simulation trials. The formula in cell B21, =B20*(B4−B18), computes the return from the auction; if Land Shark wins the auction, the return is equal to the estimated value minus the bid amount, otherwise the return is zero because the value of cell B20 will be zero.

Generating Values for Land Shark's Random Variables

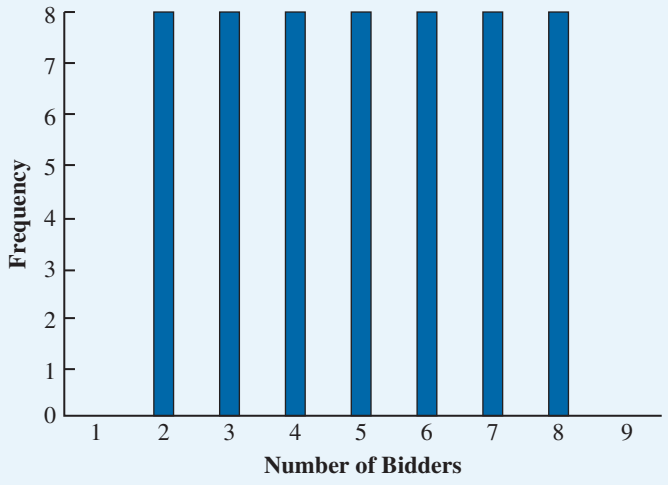
In the Land Shark simulation model constructed in Figure 13.17, the uncertain quantities are the number of competing bidders and how much the competitors will bid (as a fraction of the property's estimated value). In this section, we discuss how to specify probability distributions for these uncertain quantities, or random variables.

In Chapter 2, we discuss frequency distributions in more detail.

First, consider the number of bidders. Figure 13.18 contains the frequency distribution of the number of bidders for the 56 previous auctions that Land Shark has tracked in the *Auctions* worksheet of the file *landsharkblank*. The number of bidders has ranged from two to eight over the past 56 auctions. Unless Land Shark has reason to believe that there may be fewer than two bids on an upcoming auction, it is probably safe to assume that there will be a minimum of two competing bids. There has not been an auction with more than eight bidders, so eight is a reasonable assumption for the maximum number of competing bids unless Land Shark's experience with the local real estate market suggests that more than eight competing bids is possible.

Figure 13.18 suggests that the relative likelihood of different values for the number of bidders appears to be equal. Thus, Land Shark decides to model the number of bidders to be 2, 3, 4, 5, 6, 7, or 8 with equal probability. In this case, the integer uniform distribution is the appropriate choice, as it is characterized by a series of equally likely consecutive integers over a specified range.

Figure 13.18 Frequency Distribution of Number of Bidders in 56 Previous Auctions



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The integer uniform distribution is a special case of the discrete uniform distribution discussed in Chapter 5. In both distributions, all values are equally likely. However, in the integer uniform distribution, the possible values are consecutive integers over the defined range. In a general discrete uniform distribution, the possible values do not have to be consecutive integers (or even integers), but rather just a set of distinct, discrete values.

To generate a value for a random variable characterized by an integer uniform distribution, the following Excel formula is used:

Value of integer uniform variable
=RANDBETWEEN(lower integer value, upper integer value) **(13.8)**

For Land Shark, the lower integer value is 2 and the upper integer value is 8. Applying equation (13.8), we enter the formula =*RANDBETWEEN(2, 8)* into cell B5 as shown in Figure 13.20.

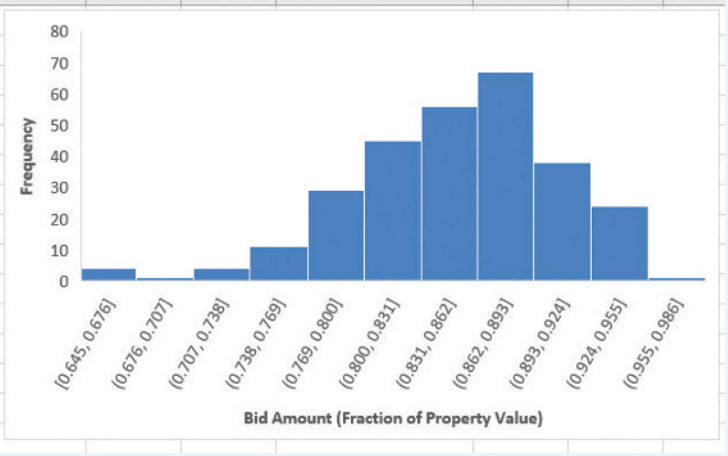
Each competitor's bid fraction is also a random variable. From the past 56 auctions, there has been a total of 280 observations of how competitors have bid (as a fraction of the respective property's estimated value). These 280 bid amounts from the *Auctions* worksheet have been relisted in the *BidList* worksheet in the file *landsharkblank*. Figure 13.19 displays the first 14 rows in the indexed list of bid fraction amounts in the *BidList* worksheet and a histogram of the bid amount data. We see that the bid amount distribution is negatively skewed, and that bid amounts most commonly occur in the range (0.862, 0.893).

There are several ways we could use the 280 bid amount observations as a basis for simulating bid amount values in our spreadsheet model. One way would be to use Figure 13.19 as the basis for choosing a discrete probability distribution to represent this uncertain value (in the same manner we generated values for direct labor cost per unit in the Sanotronics problem). However, such a discrete probability distribution would result in a loss of information, because we would need to specify one bid percentage value to represent each histogram bin. For example, if we represented the 11 histogram bins using the upper limit of each bin, then the set of 11 values {0.676, 0.707, 0.738, 0.769, 0.800, 0.831, 0.862, 0.893, 0.924, 0.955, and 0.986} would be the only possible bids in the simulation model. From the 280 observations, we see that bid percentages take on many values between the minimum of 0.645 and the maximum of 0.986. Therefore, assuming a discrete probability distribution may not be preferred for generating bid percentage values.

Two other primary alternatives are to either directly sample from the 280 observations to generate values for simulation trials, or to fit a continuous probability distribution based on the 280 observations. We will describe the approach of directly sampling from the data and discuss distribution fitting later in this section.

Figure 13.19 *BidList* Worksheet Displaying a Portion of the Indexed List of 280 Bid Amounts in 56 Previous Auctions and Associated Frequency Distribution

	A	B
1	Bid	Amount
2	1	0.830
3	2	0.797
4	3	0.833
5	4	0.878
6	5	0.839
7	6	0.843
8	7	0.835
9	8	0.823
10	9	0.781
11	10	0.892
12	11	0.767
13	12	0.787
14	13	0.763
15	14	0.862



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Directly sampling from data is a good modeling choice if Land Shark believes that these 280 bid fraction values are an accurate representation of the distribution of future bids. We will simulate the bids for the upcoming auction by randomly selecting a value from one of these 280 bid fraction values. To generate a bid fraction for the first competing bidder by sampling from the set of 280 past bid fractions, we use the Excel formula in cell B8 (as shown in Figure 13.20):

$$=XLOOKUP(D8,BidList!A2:A281,BidList!B2:B281,"error",0,1)$$

The Excel function =RANDBETWEEN(1,280) in cell D8 generates the index of one of the 280 past bid fraction values in the *BidList* worksheet. The XLOOKUP function then looks up the randomly generated index from cell array A2:A281 in the *BidList* worksheet and returns the bid fraction value from cell array B2:B281 in the *BidList* worksheet corresponding to the randomly generated index. For example, if the randomly generated index in cell D8 is 7, then the XLOOKUP formula in cell B8 of the *Model* worksheet would return the value 0.835, the bid fraction amount corresponding to Bid 7 in the *BidList* worksheet as shown in Figure 13.19.

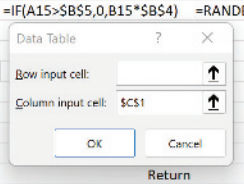
When sampling values directly from sample data, we note that only values that exist in the data will be possible values for a simulation trial. Resampling empirical data is a good approach only when the data adequately represent the range of possible values and the distribution of values across this range. If the sample data do not adequately describe the set of possible values for a random variable, it may be more appropriate to identify a probability distribution that closely fits the data and sample from the fitted probability distribution rather than just sampling directly from the data.

Executing Simulation Trials and Analyzing Output

Each trial in the simulation of the auction involves randomly generating values for the number of bidders and the eight possible bid fractions and then computing whether Land Shark wins the auction and its return from the auction. To prepare the spreadsheet for the execution of

Figure 13.20 Setting Up Land Shark Spreadsheet for 1,000 Simulation Trials

	A	B	C	D
1	**Land Shark**			
2				
3	**Parameters**			
4	Estimated Value	1389000		
5	Number of Bidders	=RANDBETWEEN(2,8)		
6				
7	Bid Index	Bid Fraction	Bid Amount	Random Sample
8	1	=XLOOKUP(D8,BidList!$A$2:$A$281,BidList!$B$2:$B$281,"error",0,1)	=IF(A8>$B$5,0,B8*$B$4)	=RANDBETWEEN(1,280)
9	2	=XLOOKUP(D9,BidList!$A$2:$A$281,BidList!$B$2:$B$281,"error",0,1)	=IF(A9>$B$5,0,B9*$B$4)	=RANDBETWEEN(1,280)
10	3	=XLOOKUP(D10,BidList!$A$2:$A$281,BidList!$B$2:$B$281,"error",0,1)	=IF(A10>$B$5,0,B10*$B$4)	=RANDBETWEEN(1,280)
11	4	=XLOOKUP(D11,BidList!$A$2:$A$281,BidList!$B$2:$B$281,"error",0,1)	=IF(A11>$B$5,0,B11*$B$4)	=RANDBETWEEN(1,280)
12	5	=XLOOKUP(D12,BidList!$A$2:$A$281,BidList!$B$2:$B$281,"error",0,1)	=IF(A12>$B$5,0,B12*$B$4)	=RANDBETWEEN(1,280)
13	6	=XLOOKUP(D13,BidList!$A$2:$A$281,BidList!$B$2:$B$281,"error",0,1)	=IF(A13>$B$5,0,B13*$B$4)	=RANDBETWEEN(1,280)
14	7	=XLOOKUP(D14,BidList!$A$2:$A$281,BidList!$B$2:$B$281,"error",0,1)	=IF(A14>$B$5,0,B14*$B$4)	=RANDBETWEEN(1,280)
15	8	=XLOOKUP(D15,BidList!$A$2:$A$281,BidList!$B$2:$B$281,"error",0,1)	=IF(A15>$B$5,0,B15*$B$4)	=RANDBETWEEN(1,280)
16				
17	**Model**			
18	Land Shark Bid Amount	1229000		
19	Largest Competitor Bid	=MAX(C8:C15)		
20	Land Shark Win Auction?	=IF(B18>B19,1,0)		
21	Land Shark Return	=B20*(B4-B18)		
22				
23	Simulation Trial		Win?	Return
24	1	=B20	=B21	
25	2			
26	3			
27	4			
28	5			



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

1,000 simulation trials, we structure the spreadsheet as in Figure 13.20. We construct a skeleton of a table in cells A23 through C1023 to hold the set of 1,000 simulation trials. Cell range A24:A1023 numbers the rows that will correspond to the 1,000 simulation trials (only the first five rows are displayed in Figure 13.20). In the first row of the table cells B24 and C24 contain Excel formulas referencing the two output measures (whether Land Shark wins the auction and its return from the auction).

To populate the table of simulation trials in the *Model* worksheet, we execute the following steps:

Step 1. Select cell range A24:L1023
Step 2. Click the **Data** tab in the Ribbon
Step 3. Click **What-If Analysis** in the **Forecast** group and select **Data Table…**
Step 4. When the **Data Table** dialog box appears, leave the **Row input cell:** box blank and enter any empty cell in the spreadsheet (e.g., *C1*) into the **Column input cell:** box
Step 5. Click **OK**

Figure 13.21 shows the results of a set of 1,000 simulation trials. After executing the simulation with the Data Table, each row of this table corresponds to a distinct simulation trial consisting of different values of the random variables. We see that Land Shark does not win the simulated auction corresponding to Trial 1, but does win the simulated auction in Trial 11 and earns a return of $160,000.

Similar to the examples in the previous sections, we compute sample statistics, 95% confidence intervals, and a 95% probability interval, based on the output of the 1,000 simulation trials. In Figure 13.21, this analysis resides in cell range E24:F40. In addition, we construct a histogram on the 1,000 values of Land Shark's return. The formulas of the summary statistics in cell range F24:F40 are analogous to the Sanotronics example with the exception of the calculation of the sample proportion in cell F36. In cell F36, the probability of Land Shark winning the auction is computed using the values of the binary outcome value in the "Win?" column with the following formula:

Cell F36 = *SUM(B24:B1023)/F24*

MODEL file
landsharkresample

Figure 13.21 shows that based on this set of 1,000 simulation trials, Land Shark's estimated mean return is $37,440 and a 95% confidence interval on the mean return is $33,234 to $41,646. The estimated probability that it wins the auction is 0.234 with an associated 95% confidence interval of 0.208 to 0.260. As the histogram in Figure 13.21 shows, when Land Shark bids $1,229,000, there are only two outcomes: either it wins the auction and earns a return of $160,000 or it loses the auction and earns a return of $0.

We note that a different set of 1,000 simulation trials can be generated by pressing the **F9** key, and this will result in different values of the summary statistics because these will now be based on a different sample. By pressing the **F9** key and observing how much the output statistics vary, the analyst can gauge how much sampling error exists in the output statistics. Furthermore, the 95% confidence interval on the mean return and the 95% confidence interval on the probability of winning the auction reflect the degree of the sampling error. Wider confidence intervals reflect more uncertainty in the accuracy of the sample mean and sample proportion. If we would press the F9 key 100 times to create 100 different samples of 1,000 trials, we would expect 95 of the corresponding 100 confidence intervals on the mean return to contain Land Shark's true mean return from the auction. Similarly, we would expect 95 of the 100 confidence intervals on the proportion of auction trials that Land Shark wins to contain the true probability of Land Shark winning the auction.

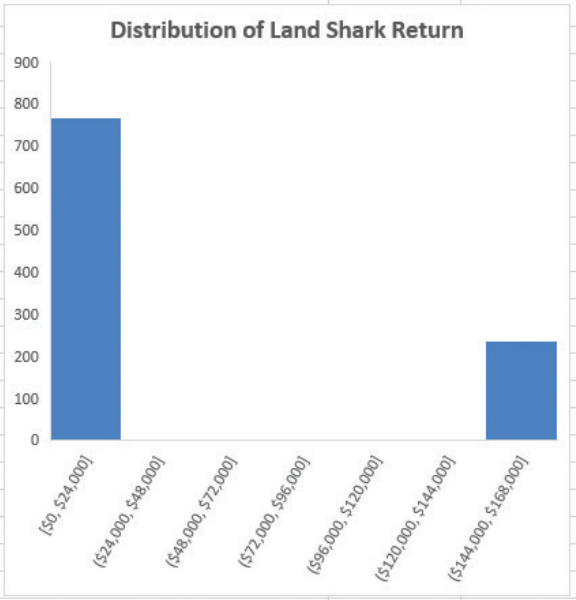
In general, increasing the number of trials in a simulation experiment will decrease the variability in the summary statistics from one sample of simulation trials to the next. Therefore, if we wish to decrease the sampling error in the output statistics, we should increase the number of simulation trials and re-execute the simulation experiment.

*When you run your LandShark simulation, the values you see will be different. This is to be expected with simulation models. Each time the simulation is executed, the values may vary because different random numbers are being used. If a set of static values is desired, you can replace the dynamic table with a static set of trial values using the Excel functionality to **Copy** and **Paste Values**.*

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.21 Output from Land Shark Simulation Based on Resampling Bid Fraction Data

	A	B	C	D
1	**Land Shark**			
2				
3	**Parameters**			
4	Estimated Value	$1,389,000		
5	Number of Bidders	7		
6				
7	Bid Index	Bid Fraction	Bid Amount	Random Sample
8	1	0.860	$1,194,540	139
9	2	0.884	$1,227,876	219
10	3	0.893	$1,240,377	158
11	4	0.882	$1,225,098	270
12	5	0.787	$1,093,143	12
13	6	0.944	$1,311,216	126
14	7	0.882	$1,225,098	191
15	8	0.841	$0	104
16				
17	**Model**			
18	Land Shark Bid Amount	$1,229,000		
19	Largest Competitor Bid	$1,311,216		
20	Land Shark Win Auction?	0		
21	Land Shark Return	$0		
22				
23	Simulation Trial	Win?	Return	
24	1	0	$0	
25	2	0	$0	
26	3	0	$0	
27	4	0	$0	
28	5	0	$0	
29	6	0	$0	
30	7	0	$0	
31	8	0	$0	
32	9	0	$0	
33	10	0	$0	
34	11	1	$160,000	
35	12	0	$0	
36	13	0	$0	
37	14	0	$0	
38	15	0	$0	
39	16	1	$160,000	
40	17	0	$0	



Land Shark Return Summary Statistics	
Count	1000
Minimum	$0
Maximum	$160,000
Mean	$37,440
Standard Deviation	$67,773
95% C.I. on Mean Return	$33,234
	$41,646
95% P.I. on Return	$0
	$160,000
P(Win)	0.234
Standard Error of Proportion	0.013
95% C.I. on P(Win)	0.208
	0.260

Generating Bid Amounts with Fitted Distributions

Specialized simulation software such as @RISK, Crystal Ball, and Analytic Solver provide automated distribution fitting functionality.

In the Land Shark model displayed in Figure 13.21, we generated the competing bid fractions by directly sampling from the 280 bids submitted in 56 previous auctions. The advantage of this approach is that it is relatively easy to execute, but if the 280 observations do not adequately represent the possible bid fractions for the upcoming auction, then our model may not accurately represent the future auction and Land Shark's assessment of its bid amount.

In this section, we examine another approach for using the 280 bid observations to generate bid fraction values in a simulation model. Specifically, we will use the 280 bid observations to fit a continuous probability distribution to a histogram based on the data. The advantage of fitting a distribution is that it will allow us to generate values that may

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

not exist in the list of the original 280 observations, but still share characteristics with these data. The disadvantage of fitting a distribution is that the process is a bit more involved and requires more familiarity with probability distributions.

Our goal is to identify a continuous probability distribution that fits the histogram of the bid fraction data shown in Figure 13.19. Appendix 13.1 contains a description of several continuous and discrete probability distributions. For the bid fraction data, we seek a *continuous* probability distribution due to the large number of possible values for a submitted bid fraction. Furthermore, we know that the range of bid fractions has a lower bound of zero and upper bound of one; a competitor cannot bid a negative fraction and a competitor will never bid more than the property's estimated value. There are many possible continuous probability distributions that have both lower and upper bounds, but some of the most common are the uniform, triangular, and beta distributions. We will consider each of these.

In Chapter 5, we discuss probability distributions in more detail.

The uniform distribution assumes each value between a specified minimum value and minimum value is *equally likely*, which does not appear to be the case for bid fractions as illustrated by Figure 13.19. So, the uniform distribution does not appear to be a good choice to generate bid fraction values. Nonetheless, if we wanted to use a uniform distribution to generate bid fractions in our simulation model, we only need to determine the minimum and maximum values. For these data, the minimum is 0.645 and the maximum is 0.986, but, theoretically, bid fractions could extend from 0.000 to 1.000. Setting the minimum and maximum of the distribution is a modeling choice that will affect how low and high our competitors will bid in the simulated auctions. If Land Shark believes that the observed values of 0.645 and 0.986 are likely to be the lowest and highest bid amounts placed by competitors, then 0.645 and 0.986 should be used as the lower and upper limits of a uniform distribution. To generate a value from a continuous uniform distribution in Excel, we can use equation (13.3) as we did in the Sanotronics problem.

The triangular distribution is a unimodal distribution characterized by three input parameters: minimum (a), mode (m), and maximum (b). While the shape of the bid fraction distribution does not appear exactly triangular, it could be worthwhile option to explore. To determine the mode (most likely) value of the triangular distribution, we note that computing the mode of the effectively continuous bid fraction data is a bit dubious as no single value occurs frequently. Therefore, we base the mode on the histogram in Figure 13.19. We observe the most frequent bin is [0.862, 0.893) and we could use the midpoint of this bin, 0.8775, as the mode of the triangular distribution. In Figure 13.22, the orange triangle-shaped curve represents the (continuous) triangular distribution with $a = 0.645$ (minimum), $m = 0.8775$ (mode), and $b = 0.986$ (maximum).

MODEL file
triangular

To generate a value for a random variable characterized by a triangular distribution, the following Excel formula is used:

Value of triangular random variable

$$=\text{IF}(random < (m-a)/(b-a),\ a + \text{SQRT}((b-a)*(m-a)*random),\ b - \text{SQRT}((b-a)*(b-m)*(1-random))) \quad \textbf{(13.9)}$$

The file triangular *implements equation (13.9) to facilitate the fitting of a triangular distribution to data.*

In equation (13.9), *random* refers to a single, separate cell containing the Excel function =RAND(); a single, separate cell is necessary to make sure the same random value is used everywhere *random* appears in equation (13.9). Applying equation (13.9) for the triangular distribution fit to the 280 bid observations yields:

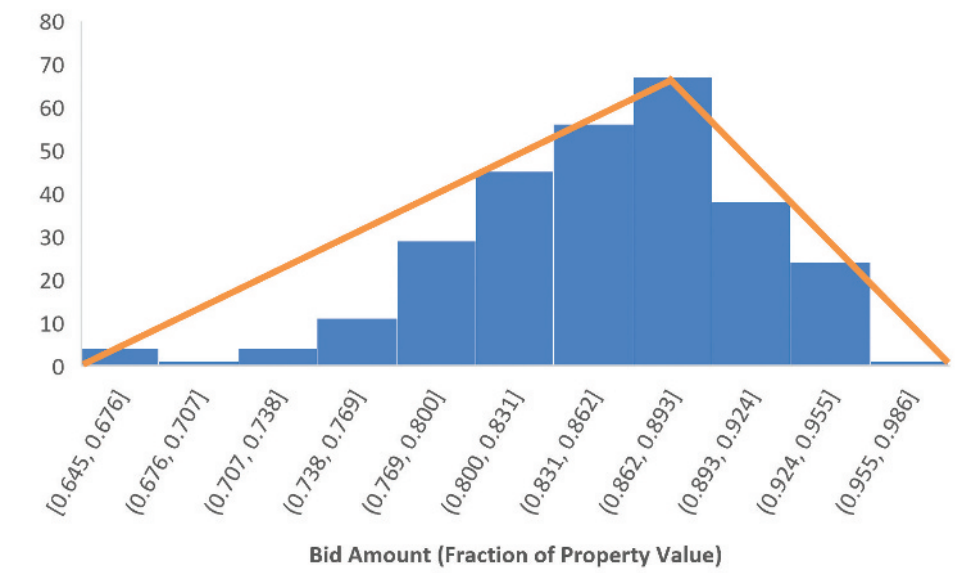
$$\text{Bid fraction} = \text{IF}(random < (0.8775-0.645)/(0.986-0.645),\ 0.645 + \text{SQRT}((0.986-0.645)*(0.8775-0.645)*random),\ 0.986 - \text{SQRT}((0.986-0.645)*(0.986-0.8775)*(1-random))) \quad \textbf{(13.10)}$$

MODEL file
landsharktriangular

In the file *landsharktriangular*, the worksheet *TriangularFit* contains the Excel implementation of fitting the triangular distribution to the 280 bid fraction observations and generating values from this triangular distribution. As Figure 13.23 shows, sample statistics for

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.22 Fit of Triangular Distribution to Bid Fraction Data



the 280 bid fractions are computed in the cell range E2:E6 of the *TriangularFit* worksheet. Cells E8, E9, and E10 perform the preliminary calculation of values used in the implementation of equation (13.9). The cell range D12:F20 contains a table of the eight values generated from the fitted triangular distribution. For example, cell E13 implements equation (13.9) using the random number from the =RAND() function in cell F13 to generate the first value from the triangular distribution in the table.

Now that the worksheet *TriangularFit* in the file *landsharktriangular* randomly generates eight bid fraction values from a fitted triangular distribution, we can alter the Land Shark simulation model to incorporate these values. As Figure 13.24 shows, the eight bid fraction values in the cell range B8:B15 of the *Model* worksheet are generated by referencing the *TriangularFit* worksheet. For example, the first competitor bid fraction in cell B8 is now computed by =TriangularFit!E13, which references the first value generated from the fitted triangular distribution in worksheet *TriangularFit.* From Figure 13.25, we see that modeling bid fraction values with a triangular distribution results in a 95% confidence interval of $34,004 to $42,476 on the mean return and a 95% confidence interval of 0.213 to 0.265 on the probability of winning the auction.

By comparing the simulation output in Figure 13.25 to the simulation output in Figure 13.21 based on generating bid fraction values by directly sampling the past 280 bid observations, we can gauge the effect of changing how we model the uncertain competitor bidding. The results in Figure 13.25 using the triangular distribution to generate competitor bid amounts are slightly more optimistic than the results in Figure 13.21. However, the slight differences in the simulation output of Figures 13.25 and 13.21 do not appear statistically significant based on 1,000 simulation trials.

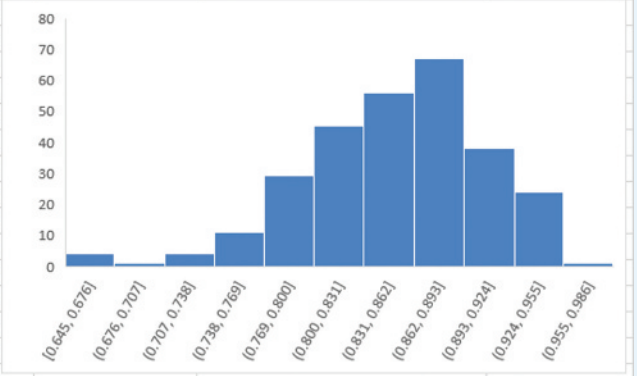
The final alternative for modeling the bid fraction values would be to fit a beta distribution to the 280 bid observations. The beta distribution is a very flexible distribution characterized by four input parameters: alpha (α), beta (β), minimum (A), and maximum (B). A common method for estimating the α and β values in a beta distribution uses the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.23 Generating Values from a Triangular Distribution in *TriangularFit* Worksheet

	A	B	C	D	E	F
1	Bid	Amount		Sample Statistics		
2	1	0.83		minimum	=MIN(B:B)	
3	2	0.797		maximum	=MAX(B:B)	
4	3	0.833		mean	=AVERAGE(B:B)	
5	4	0.878		standard deviation	=STDEV.S(B:B)	
6	5	0.839		mode	=(0.862+0.893)/2	
7	6	0.843				
8	7	0.835		Max - Min	=E3-E2	
9	8	0.823		Mode - Min	=E6-E2	
10	9	0.781		Max - Mode	=E3-E6	
11	10	0.892				
12	11	0.767		Observation	Value from Triangular Distribution	Random Number
13	12	0.787		1	=IF(F13<$E$9/$E$8,$E$2+SQRT($E$8*$E$9*F13),$E$3-SQRT($E$8*$E$10*(1-F13)))	=RAND()
14	13	0.763		2	=IF(F14<$E$9/$E$8,$E$2+SQRT($E$8*$E$9*F14),$E$3-SQRT($E$8*$E$10*(1-F14)))	=RAND()
15	14	0.862		3	=IF(F15<$E$9/$E$8,$E$2+SQRT($E$8*$E$9*F15),$E$3-SQRT($E$8*$E$10*(1-F15)))	=RAND()
16	15	0.814		4	=IF(F16<$E$9/$E$8,$E$2+SQRT($E$8*$E$9*F16),$E$3-SQRT($E$8*$E$10*(1-F16)))	=RAND()
17	16	0.895		5	=IF(F17<$E$9/$E$8,$E$2+SQRT($E$8*$E$9*F17),$E$3-SQRT($E$8*$E$10*(1-F17)))	=RAND()
18	17	0.771		6	=IF(F18<$E$9/$E$8,$E$2+SQRT($E$8*$E$9*F18),$E$3-SQRT($E$8*$E$10*(1-F18)))	=RAND()
19	18	0.859		7	=IF(F19<$E$9/$E$8,$E$2+SQRT($E$8*$E$9*F19),$E$3-SQRT($E$8*$E$10*(1-F19)))	=RAND()
20	19	0.867		8	=IF(F20<$E$9/$E$8,$E$2+SQRT($E$8*$E$9*F20),$E$3-SQRT($E$8*$E$10*(1-F20)))	=RAND()
21	20	0.85				
22	21	0.833				
23	22	0.836				
24	23	0.898				
25	24	0.831				
26	25	0.897				
27	26	0.831				
28	27	0.657				
29	28	0.846				
30	29	0.85				
31	30	0.863				
32	31	0.825				
33	32	0.91				
34	33	0.848				
35	34	0.89				

	A	B	C	D	E	F
1	**Bid**	**Amount**		Sample Statistics		
2	1	0.830		minimum	0.645	
3	2	0.797		maximum	0.986	
4	3	0.833		mean	0.852	
5	4	0.878		standard deviation	0.057	
6	5	0.839		mode	0.878	
7	6	0.843				
8	7	0.835		Max - Min	0.341	
9	8	0.823		Mode - Min	0.233	
10	9	0.781		Max - Mode	0.109	
11	10	0.892				
12	11	0.767		Observation	Value from Triangular Distribution	Random Number
13	12	0.787		1	0.893	0.7648
14	13	0.763		2	0.884	0.7163
15	14	0.862		3	0.860	0.5837
16	15	0.814		4	0.882	0.7062
17	16	0.895		5	0.925	0.8988
18	17	0.771		6	0.875	0.6648
19	18	0.859		7	0.944	0.9525
20	19	0.867		8	0.845	0.5029
21	20	0.85				
22	21	0.833				
23	22	0.836				
24	23	0.898				
25	24	0.831				
26	25	0.897				
27	26	0.831				
28	27	0.657				
29	28	0.846				
30	29	0.850				
31	30	0.863				
32	31	0.825				
33	32	0.91				
34	33	0.848				
35	34	0.89				
36	35	0.986				



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.24 Land Shark Formula Worksheet for Bid Fraction Value Generated from Triangular Distribution

	A	B	C
1	**Land Shark**		
2			
3	**Parameters**		
4	Estimated Value	1389000	
5	Number of Bidders	=RANDBETWEEN(2,8)	
6			
7	Bid Index	Bid Fraction	Bid Amount
8	1	=TriangularFit!E13	=IF(A8>$B$5,0,B8*$B$4)
9	2	=TriangularFit!E14	=IF(A9>$B$5,0,B9*$B$4)
10	3	=TriangularFit!E15	=IF(A10>$B$5,0,B10*$B$4)
11	4	=TriangularFit!E16	=IF(A11>$B$5,0,B11*$B$4)
12	5	=TriangularFit!E17	=IF(A12>$B$5,0,B12*$B$4)
13	6	=TriangularFit!E18	=IF(A13>$B$5,0,B13*$B$4)
14	7	=TriangularFit!E19	=IF(A14>$B$5,0,B14*$B$4)
15	8	=TriangularFit!E20	=IF(A15>$B$5,0,B15*$B$4)
16			
17	**Model**		
18	Land Shark Bid Amount	1229000	
19	Largest Competitor Bid	=MAX(C8:C15)	
20	Land Shark Win Auction?	=IF(B18>B19,1,0)	
21	Land Shark Return	=B20*(B4-B18)	
22	Land Shark Return ($1000)	=B21/1000	

MODEL *file*
beta

The file beta *implements equations (13.11) and (13.12) to facilitate the fitting of a beta distribution to data.*

sample mean ($\bar{x}$) and sample standard deviation (s) as shown in equations (13.11) and (13.12) below.[6]

$$\alpha = \left(\frac{\bar{x}-A}{B-A}\right)\left(\frac{\left(\frac{\bar{x}-A}{B-A}\right)\left(1-\left(\frac{\bar{x}-A}{B-A}\right)\right)}{\frac{s^2}{(B-A)^2}}-1\right) \tag{13.11}$$

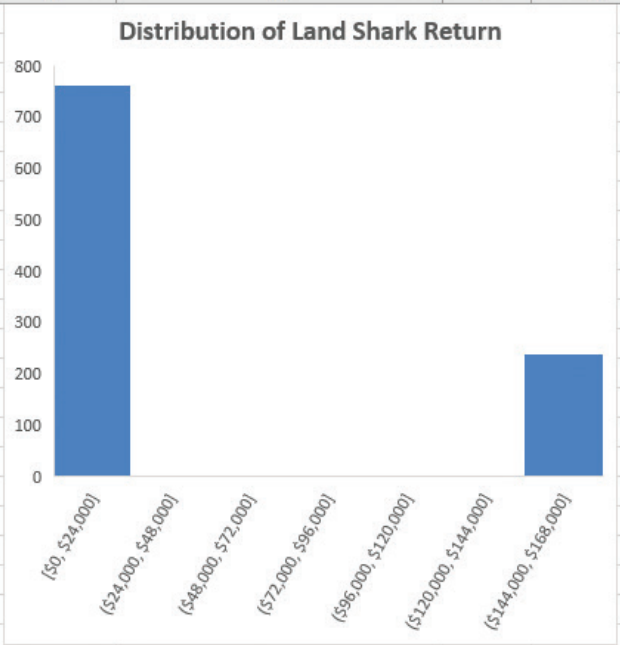
$$\beta = \alpha \times \left(\frac{\left(1-\left(\frac{\bar{x}-A}{B-A}\right)\right)}{\left(\frac{\bar{x}-A}{B-A}\right)}\right) \tag{13.12}$$

[6]Estimating the parameters using equations (13.11) and (13.12) is based on a statistical method known as the "method of moments." The specifics of this method are beyond the scope of this textbook.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.25 Output from Land Shark Simulation Using Triangular Distribution to Generate Bid Fraction Values

	A	B	C
1	**Land Shark**		
2			
3	**Parameters**		
4	Estimated Value	$1,389,000	
5	Number of Bidders	6	
6			
7	Bid Index	Bid Fraction	Bid Amount
8	1	0.838	$1,164,429
9	2	0.852	$1,183,555
10	3	0.738	$1,025,556
11	4	0.838	$1,164,030
12	5	0.928	$1,288,992
13	6	0.834	$1,158,779
14	7	0.760	$0
15	8	0.834	$0
16			
17	**Model**		
18	Land Shark Bid Amount	$1,229,000	
19	Largest Competitor Bid	$1,288,992	
20	Land Shark Win Auction?	0	
21	Land Shark Return	$0	
22			
23	Simulation Trial	Win?	Return ($1000)
24	1	0	$0
25	2	0	$0
26	3	0	$0
27	4	0	$0
28	5	1	$160,000
29	6	1	$160,000
30	7	0	$0
31	8	0	$0
32	9	0	$0
33	10	0	$0
34	11	0	$0
35	12	0	$0
36	13	1	$160,000
37	14	1	$160,000
38	15	0	$0
39	16	0	$0
40	17	0	$0



Land Shark Return Summary Statistics	
Count	1000
Minimum	$0
Maximum	$160,000
Mean	$38,240
Standard Deviation	$68,270
95% C.I. on Mean Return	$34,004
	$42,476
95% P.I. on Return	$0
	$160,000
P(Win)	0.239
Standard Error of Proportion	0.013
95% C.I. on P(Win)	0.213
	0.265

For the 280 bid observations, the sample mean is $\bar{x} = 0.852$, the sample standard deviation is $s = 0.057$, the minimum value is 0.645, and the maximum value is 0.986. Substituting these values first into equation (13.11) and then into equation (13.12) provides:

$$\alpha = \left(\frac{0.852 - 0.645}{0.986 - 0.645}\right)\left(\frac{\frac{0.852 - 0.645}{0.986 - 0.645}\left(1 - \left(\frac{0.852 - 0.645}{0.986 - 0.645}\right)\right)}{\frac{0.057^2}{(0.986 - 0.645)^2}} - 1\right) = 4.566$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$$\beta = 4.566 \times \left(\frac{\left(1 - \left(\frac{0.852 - 0.645}{0.986 - 0.645}\right)\right)}{\left(\frac{0.852 - 0.645}{0.986 - 0.645}\right)} \right) = 2.974$$

To generate a value for a random variable characterized by a beta distribution, the following Excel formula is used:

Value of beta random variable

$$=\text{BETA.INV}(\text{RAND}(), \alpha, \beta, A, B) \quad \textbf{(13.13)}$$

For the Land Shark problem, substituting the values of the parameters results in:

$$\text{Bid fraction} = \text{BETA.INV}(\text{RAND}(), 4.566, 2.974, 0.645, 0.986) \quad \textbf{(13.14)}$$

Figure 13.26 provides a visualization of the beta distribution's fit to the bid fraction data. In Figure 13.26, the orange curve represents the (continuous) beta distribution with $\alpha = 4.566$, $\beta = 2.974$, $A = 0.645$ (minimum), and $B = 0.986$ (maximum).

Figure 13.27 displays the formula view of the Land Shark simulation model implementing equation (13.14) to generate bid fraction values. From Figure 13.28, we see that modeling bid fraction values with a beta distribution results in a 95% confidence interval of \$27,569 to \$35,471 on the mean return and a 95% confidence interval of 0.172 to 0.222 on the probability of winning the auction.

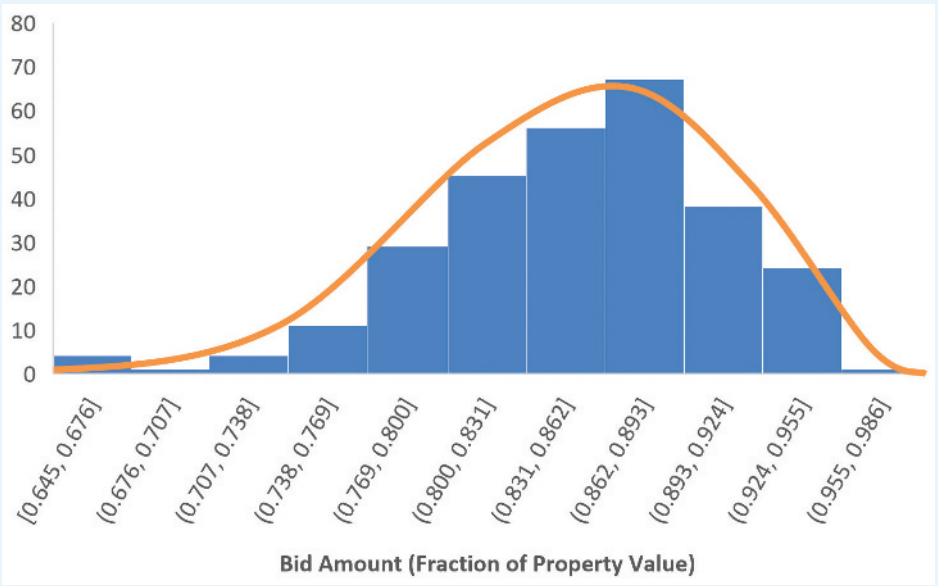
By comparing the simulation output in Figure 13.28 to the simulation output in Figures 13.21 and 13.25, we can gauge the effect of using the beta distribution to generate competitor bid amounts versus directly sampling or the triangular distribution, respectively. The results in Figure 13.28 are the least optimistic of these three approaches. However, the slight differences in the simulation output of Figures 13.21, 13.25, and 13.28 do not appear statistically significant based on 1,000 simulation trials.

While it is impossible to discern what is the "best" way to model the uncertain bid fraction values, the exercise of testing different distributions generates insight. One benefit of using a good-fitting theoretical distribution (such as the beta distribution in this case) to generate bid fraction values is that it generates thousands of unique bid fractions and is not limited to bid fraction values seen in the past. Conversely, sampling directly from the observed data means that the only values that can be used as inputs into the simulation are those 280 bid fraction values that were observed in the past and that values may get reused multiple times.

In general, the appropriate way to generate values for the random variables in a Monte Carlo simulation may be difficult to determine. For a well-defined situation, like rolling a fair die, it may be clear how to generate the value of the random variable (the outcome of a dice roll). In other situations, we may not know exactly how to model the uncertainty. In these situations, it is recommended that we examine any sample data available to us. The sample data can then be used by sampling directly or we can compare the sample data to common probability distributions (such as uniform, normal, triangular, and beta distributions) to determine if we can approximate the distribution of the data with an existing probability distribution. In all cases, it is important to test the implications of different modeling approaches and to understand that a simulation model does not allow us to perfectly see the future, but rather it helps us to understand the impact of uncertainty on our decisions.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.26 Fit of Beta Distribution to Bid Fraction Data



MODEL file
landsharkbeta

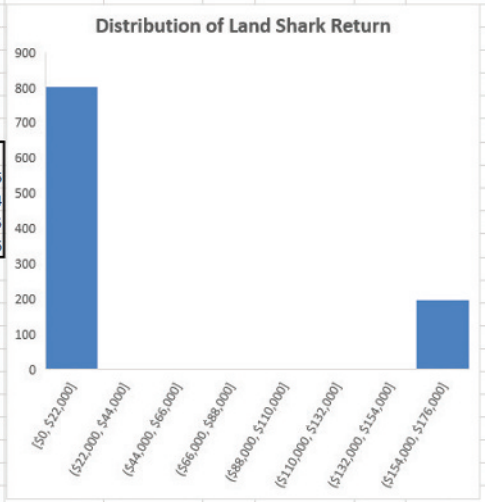
Figure 13.27 Land Shark Formula Worksheet for Bid Fraction Value Generated from Beta Distribution

	A	B	C	D	E	F
1	**Land Shark**					
2						
3	**Parameters**					
4	Estimated Value	1389000				
5	Number of Bidders	=RANDBETWEEN(2,8)				
6						
7	Bid Index	Bid Fraction	Bid Amount			
8	1	=BETA.INV(RAND(),$F$9,$F$10,$F$11,$F$12)	=IF(A8>$B$5,0,B8*$B$4)		Bid % Parameters (Beta)	
9	2	=BETA.INV(RAND(),$F$9,$F$10,$F$11,$F$12)	=IF(A9>$B$5,0,B9*$B$4)		alpha	4.56599625463373
10	3	=BETA.INV(RAND(),$F$9,$F$10,$F$11,$F$12)	=IF(A10>$B$5,0,B10*$B$4)		beta	2.97358699757436
11	4	=BETA.INV(RAND(),$F$9,$F$10,$F$11,$F$12)	=IF(A11>$B$5,0,B11*$B$4)		minimum	0.645
12	5	=BETA.INV(RAND(),$F$9,$F$10,$F$11,$F$12)	=IF(A12>$B$5,0,B12*$B$4)		maximum	0.986
13	6	=BETA.INV(RAND(),$F$9,$F$10,$F$11,$F$12)	=IF(A13>$B$5,0,B13*$B$4)			
14	7	=BETA.INV(RAND(),$F$9,$F$10,$F$11,$F$12)	=IF(A14>$B$5,0,B14*$B$4)			
15	8	=BETA.INV(RAND(),$F$9,$F$10,$F$11,$F$12)	=IF(A15>$B$5,0,B15*$B$4)			
16						
17	**Model**					
18	Land Shark Bid Amount	1229000				
19	Largest Competitor Bid	=MAX(C8:C15)				
20	Land Shark Win Auction?	=IF(B18>B19,1,0)				
21	Land Shark Return	=B20*(B4-B18)				

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.28 Output from Land Shark Simulation Using Beta Distribution for Bid Fraction Values

	A	B	C	D	E	F
1	**Land Shark**					
2						
3	**Parameters**					
4	Estimated Value	$1,389,000				
5	Number of Bidders	6				
6						
7	Bid Index	Bid Fraction	Bid Amount			
8	1	0.899	$1,249,306		Bid % Parameters (Beta)	
9	2	0.831	$1,154,926		alpha	4.566
10	3	0.918	$1,274,826		beta	2.974
11	4	0.917	$1,273,302		minimum	0.645
12	5	0.743	$1,032,220		maximum	0.986
13	6	0.919	$1,276,355			
14	7	0.864	$0			
15	8	0.870	$0			
16						
17	**Model**					
18	Land Shark Bid Amount	$1,229,000				
19	Largest Competitor Bid	$1,276,355				
20	Land Shark Win Auction?	0				
21	Land Shark Return	$0				
22						
23	Simulation Trial	Win?	Return		Land Shark Return Summary Statistics	
24	1	0	$0		Count	1000
25	2	0	$0		Minimum	$0
26	3	0	$0		Maximum	$160,000
27	4	0	$0		Mean	$31,520
28	5	0	$0		Standard Deviation	$63,669
29	6	1	$160,000			
30	7	0	$0		95% C.I. on Mean Return	$27,569
31	8	0	$0			$35,471
32	9	0	$0			
33	10	0	$0		95% P.I. on Return	$0
34	11	0	$0			$160,000
35	12	0	$0			
36	13	0	$0		P(Win)	0.197
37	14	1	$160,000		Standard Error of Proportion	0.013
38	15	0	$0			
39	16	0	$0		95% C.I. on P(Win)	0.172
40	17	0	$0			0.222



Notes + Comments

In the Land Shark example, the finite range of possible bid fraction values was leveraged to focus the analysis on distributions with lower and upper bounds (uniform, triangular, beta). In this way, the analyst could ensure that no values outside a specified range would be generated.

Another way to avoid generating values outside a specified range is to truncate the distribution. For example, suppose an analyst wants to model an uncertain quantity as a normal random variable, but wants to ensure that no negative values are randomly generated (negative values may occur in a normal distribution because it has no lower bound). For a normal distribution with mean m and standard deviation s, the probability of negative value is given by p = NORM.DIST(0, m, s, 1). The generation of a non-negative random value from a normal distribution with mean m and standard deviation s can then be generated by the Excel formula =NORM.INV(p+(RAND()*(1-p)),m,s). Observe that when p is zero, this formula simplifies to how we generate a value from a normal distribution without any consideration of its sign. Thus, the analyst only needs to worry about dubious negative values from a normal distribution when p is not close to zero.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

13.4 Simulation with Dependent Random Variables

In the examples of the previous sections, we generated values of each uncertain quantity independently of each other. In other words, we treated each uncertain quantity as an independent random variable. In this section, we consider an example in which the values of some of the uncertain quantities are dependent.

Press Teag Worldwide (PTW) manufactures all of its products in the United States, but it sells the items in three different overseas markets: the United Kingdom, New Zealand, and Japan. Each of these overseas markets generates revenue in a different currency: pound sterling in the United Kingdom, New Zealand dollars in New Zealand and yen in Japan. At the end of each 13-week quarter, PTW converts the revenue from these three overseas markets back into U.S. dollars in order to pay its expenses in the United States, exposing PTW to exchange rate risk.

Spreadsheet Model for Press Teag Worldwide

To assess the degree of PTW's exposure to quarterly fluctuations in exchange rates, we develop a simulation model. The first step is to identify the input parameters and output measures. The next step is to develop a spreadsheet model that computes the values of the output measures given value of the input parameters. Then we prepare the spreadsheet model for simulation analysis by replacing the static values of the input parameters that are uncertain with probability distributions of possible values.

The relevant input parameters are: (i) the quarterly revenue generated in each of the three foreign currencies and (ii) the end-of-quarter exchange rates between these foreign currencies and the U.S. dollar. The output measure of interest is the total end-of-quarter revenues converted into U.S. dollars.

To model the fluctuation in the exchange rate between the pound sterling and the U.S. dollar over the next quarter, PTW expresses the number of pounds sterling (£) per U.S. dollar ($) by

$$\begin{aligned}&\text{(end-of-quarter £/\$ rate)}\\&\quad=\text{(start-of-quarter £/\$ rate)}\times(1+\%\text{ change in £/\$ rate})\end{aligned}\qquad\textbf{(13.15)}$$

That is, equation (13.15) computes the end-of-quarter exchange rate based on the start-of-quarter exchange rate and the percent change in the exchange rate over the quarter. Analogously, the equations computing the end-of-quarter exchange rates between New Zealand dollars (NZD) per U.S. dollar and Japanese yen (¥) per U.S. dollar are as follows:

$$\begin{aligned}&\text{(end-of-quarter NZD/\$ rate)}\\&\quad=\text{(start-of-quarter NZD/\$ rate)}\times(1+\%\text{ change in NZD/\$ rate})\end{aligned}\qquad\textbf{(13.16)}$$

$$\begin{aligned}&\text{(end-of-quarter ¥/\$ rate)}\\&\quad=\text{(start-of-quarter ¥/\$ rate)}\times(1+\%\text{ change in ¥/\$ rate})\end{aligned}\qquad\textbf{(13.17)}$$

To see how one would use equations (13.15), (13.16), and (13.17), suppose that the start-of-quarter exchange rates are £0.615 per U.S. dollar, NZD 1.200 per U.S. dollar, and ¥87.10 per U.S. dollar. Further, assume that there is a 4.61 percent increase in the £ per $ exchange rate, a 0.27 percent decrease in the NZD per $ exchange rate, and a 11.23 percent increase in the ¥ per $ exchange rate. Then, we would have the following:

$$\begin{aligned}\text{(end-of-quarter £/\$ rate)}&=0.615\times(1+0.0461)=\text{£}0.6436\text{ per \$}\\\text{(end-of-quarter NZD/\$ rate)}&=1.200\times(1+(-0.0027))=\text{NZD }1.1968\text{ per \$}\\\text{(end-of-quarter ¥/\$ rate)}&=87.10\times(1+0.1123)=\text{¥}96.8813\text{ per \$}\end{aligned}$$

Once the end-of-quarter exchange rates are known, the quarterly revenue in pounds sterling, New Zealand dollar, and Japanese yen can be converted into U.S. dollars as follows:

$$\begin{aligned}&\text{(end-of-quarter \$ from £)}\\&\quad=\text{(quarterly revenue in £)}\div\text{(end-of-quarter £/\$ rate)}\end{aligned}\qquad\textbf{(13.18)}$$

$$\begin{aligned}&\text{(end-of-quarter \$ from NZD)}\\&\quad=\text{(quarterly revenue in NZD)}\div\text{(end-of-quarter NZD/\$ rate)}\end{aligned}\qquad\textbf{(13.19)}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(end-of-quarter $ from ¥)

$$= \text{(quarterly revenue in ¥)} \div \text{(end-of-quarter ¥/\$ rate)} \quad \textbf{(13.20)}$$

As an illustration of these calculations, suppose the quarterly revenues generated in pounds sterling, New Zealand dollar, and the Japanese yen are £100,000, NZD 250,000, and ¥10,000,000, respectively. Then, applying equations (13.18), (13.19), and (13.20), we compute the following:

$$\text{(end-of-quarter \$ from £)} = £100{,}000 \div £0.6436 \text{ per \$} = \$155{,}385$$
$$\text{(end-of-quarter \$ from NZD)} = \text{NZD}250{,}000 \div \text{NZD } 1.1968 \text{ per \$} = \$208{,}897$$
$$\text{(end-of-quarter \$ from ¥)} = ¥10{,}000{,}000 \div ¥96.8813 \text{ per \$} = \$103{,}219$$

MODEL file

quarterlyexchangeblank

The total revenue in U.S. dollars is then $\$155{,}385 + \$208{,}897 + \$103{,}219 = \$467{,}502$ (after rounding). Figure 13.29 shows the formula view and value view of the PTW spreadsheet model for the base scenario just presented.

The percent change in the exchange rate between pairs of currencies from the start to the end of a quarter is uncertain. Therefore, PTW would like to use random variables to model the percent change in the £ per $ rate, the percent change in the NZD per $ rate, and the percent change in the ¥ per $ rate.

However, PTW realizes that there are dependencies between the exchange rate fluctuations. For example, if the U.S. dollar weakens against the pound sterling, it may

Figure 13.29 Base Spreadsheet Model for Press Teag Worldwide

	A	B	C	D	E
1	**Press Teag Worldwide**				
2					
3	**Parameters**				
4	Start-of-Quarter Exchange Rate (per $)	0.6152	1.2	87.1	
5	Quarterly % Change in Exchange Rate	0.0461	-0.0027	0.1123	
6	End-of-Quarter Exchange Rate (per $)	=B4*(1+B5)	=C4*(1+C5)	=D4*(1+D5)	
7	Quarterly Revenue	100000	250000	10000000	
8					
9	**Model**				Total
10	End-of-Quarter Revenue in $	=B7/B6	=C7/C6	=D7/D6	=SUM(B10:D10)
11				Total Revenue ($1000):	=E10/1000

	A	B	C	D	E
1	**Press Teag Worldwide**				
2					
3	**Parameters**				
4	Start-of-Quarter Exchange Rate (per $)	£0.615	NZD 1.200	¥87.10	
5	Quarterly % Change in Exchange Rate	4.61%	-0.27%	11.23%	
6	End-of-Quarter Exchange Rate (per $)	£0.6436	£1.1968	£96.8813	
7	Quarterly Revenue	£100,000	NZD 250,000	¥10,000,000	
8					
9	**Model**				Total
10	End-of-Quarter Revenue in $	$155,385	$208,897	$103,219	$467,502
11				Total Revenue ($1000):	$468

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

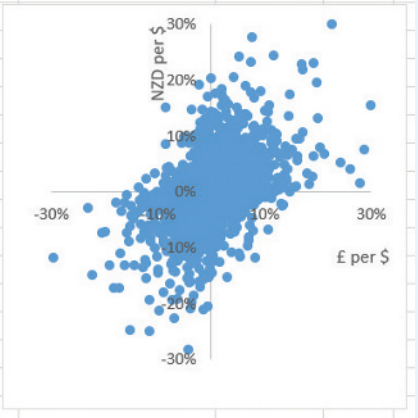
be more likely to also weaken against the New Zealand dollar. Therefore, the percent changes in the exchange rates should not be generated independently, but instead these values should be generated jointly (as a related collection of values). To account for these dependencies, PTW constructed a data set on the joint percent changes between the three exchange rates for 2,000 quarter-scenarios in the *Data* worksheet of the file *quarterlyexchange*. These data are based on historical observations as well as scenarios based on expert judgment. Figure 13.30 displays a portion of the percent change data for the three currency exchange rates.

We introduce the concept of correlation in Chapter 2.

In Figure 13.30, cell G3 contains the correlation coefficient between the quarterly percent change in £ per $ and the quarterly percent change in NZD per $ for these 2,000 quarters. This correlation of 0.478 demonstrates that there is a positive linear relationship between the quarterly percent change in £ per $ and the quarterly percent change in NZD per $. The scatter chart in Figure 13.30 corroborates the presence of this correlation. Positive percentage fluctuations of £ per $ often occur with positive percentage fluctuations of NZD per $ while negative fluctuations of £ per $ often occur with negative fluctuations of NZD per $. If these values were independent, we would expect to see no pattern in this scatter plot. Similarly, the correlation coefficient of 0.315 in cell G4 of Figure 13.30 demonstrates that there is a positive linear relationship between the quarterly percent change in £ per $ and the quarterly percent change in ¥ per $. The correlation coefficient of 0.235 in cell G5 of Figure 13.30 demonstrates that there is a positive linear relationship between the quarterly percent change in NZD per $ and the quarterly percent change in ¥ per $. While not shown in Figure 13.30, in the *Data* worksheet of the file *quarterlyexchange* there are analogous scatter charts plotting the quarterly percent change in ¥ per $ versus the quarterly percent change in £ per $, and the quarterly percent change in ¥ per $ versus the quarterly percent change in NZD per $.

Figure 13.30 *Data* Worksheet Displaying a Portion of Quarterly Percent Change Data for Three Currency Exchange Rates

	A	B	C	D	E	F	G	H
1		Quarterly Percent Changes						
2	Quarter	£ Per $	NZD per $	¥ Per $			Correlation	
3	1	-0.96%	-0.96%	-0.05%		£ per $ and NZD per $	0.478	
4	2	-0.06%	-0.05%	-0.04%		£ per $ and ¥ per $	0.315	
5	3	-3.03%	-3.02%	-7.75%		NZD per $ and ¥ per $	0.235	
6	4	-4.22%	-2.82%	-6.68%				
7	5	-0.44%	-0.08%	-1.27%				
8	6	6.64%	0.08%	-0.85%				
9	7	4.31%	0.00%	-0.03%				
10	8	-5.59%	-9.61%	-11.77%				
11	9	-3.49%	-1.05%	-1.05%				
12	10	6.52%	-1.11%	0.95%				
13	11	4.74%	-6.25%	5.52%				
14	12	-0.79%	0.35%	0.34%				
15	13	-2.47%	-1.03%	0.59%				
16	14	3.27%	1.93%	4.67%				
17	15	-1.24%	9.09%	1.65%				
18	16	-2.69%	-2.40%	-3.06%				
19	17	9.73%	4.01%	1.33%				
20	18	7.42%	23.27%	2.61%				



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

As discussed in the Land Shark example in Section 13.3, one way of generating values of a random variable is to directly sample data containing observations of this random variable. Furthermore, for the given data consisting of values of two or more random variables in the same scenario (as in the PTW data), we can capture the inherit dependency between two or more random variables by directly sampling the same scenario to generate values of these random variables. In the PTW example, to directly sample one of the 2,000 scenarios and obtain the corresponding percent change in £ per $ rate, NZD per $ rate, and ¥ per $ rate, we use the respective Excel formulas:

$$\text{Cell B5} = XLOOKUP(E5, Data!A3:A2002, Data!B3:B2002, \text{“error”}, 0, 1) \quad \textbf{(13.21)}$$

$$\text{Cell C5} = XLOOKUP(E5, Data!A3:A2002, Data!C3:C2002, \text{“error”}, 0, 1) \quad \textbf{(13.22)}$$

$$\text{Cell D5} = XLOOKUP(E5, Data!A3:A2002, Data!D3:D2002, \text{“error”}, 0, 1) \quad \textbf{(13.23)}$$

We describe the XLOOKUP function in more detail in Chapter 12.

In Figure 13.31, we demonstrate the implementation of equation (13.23) in cell D5. Columns B and C are compressed, but cells B5 and C5 contain a similar implementation of equations (13.21) and (13.22), respectively. Figure 13.31 shows that cell E5 contains the Excel function =RANDBETWEEN(1, 2000) which randomly generates the index of one of the 2,000 quarter-scenarios. The XLOOKUP function in cell D5 then looks up this index in column A of the *Data* sheet containing the quarter-scenario data and returns the corresponding percent change in ¥ per $ rate found in column D of the *Data* sheet. For example, if the randomly generated index in cell E5 is 8, then the XLOOKUP formula in cell B5 would return

Figure 13.31 Modeling Dependent Exchange Rates by Directly Resampling Data in the PTW Simulation

	A	B	C	D	E
1	**Press Teag Worldwide**				
2					
3	**Parameters**				
4	Start-of-Quarter Exchange Rate (per $)	0.6152	1.2	87.1	Random Scenario
5	Quarterly % Change in Exchange Rate	=XLOOKUP(	=XLOOKUP(	=XLOOKUP($E$5,Data!$A$3:$A$2002,Data!D3:D2002)	=RANDBETWEEN(1,2000)
6	End-of-Quarter Exchange Rate (per $)	=B4*(1+B5)	=C4*(1+C5)	=D4*(1+D5)	
7	Quarterly Revenue	100000	250000	10000000	
8					
9	**Model**				Total
10	End-of-Quarter Revenue in $	=B7/B6	=C7/C6	=D7/D6	=SUM(B10:D10)
11				Total Revenue ($1000):	=E10/1000

	A	B	C	D	E
1	**Press Teag Worldwide**				
2					
3	**Parameters**				
4	Start-of-Quarter Exchange Rate (per $)	£0.615	NZD 1.200	¥87.10	Random Scenario
5	Quarterly % Change in Exchange Rate	4.33%	2.09%	12.17%	884
6	End-of-Quarter Exchange Rate (per $)	£0.642	NZD 1.225	¥97.70	
7	Quarterly Revenue	£100,000	NZD 250,000	¥10,000,000	
8					
9	**Model**				Total
10	End-of-Quarter Revenue in $	$155,803	$204,068	$102,354	$462,225
11				Total Revenue ($1000):	$462

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the value −11.77%, the quarterly percent change in ¥ per $ corresponding to Quarter 8 in the *Data* worksheet as shown in Figure 13.30.

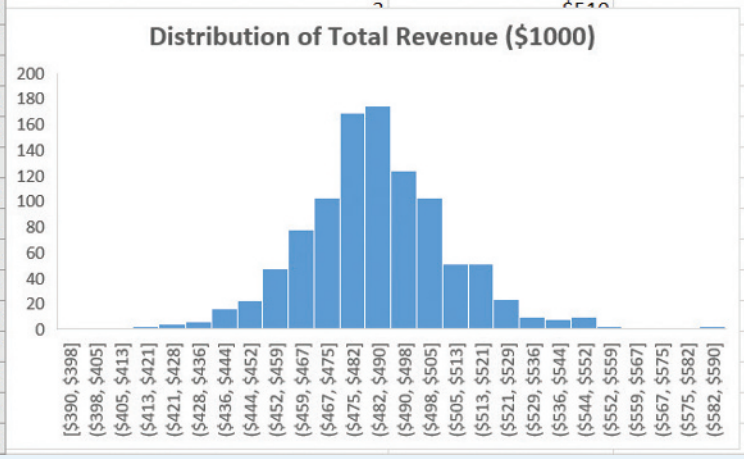
By sampling the exchange rate fluctuations computed in cells B5, C5, and D5 from the same quarter-scenario (specified by cell E5), any interdependency between these quantities is implicitly captured. That is, the individual exchange rate changes are not generated independently, but rather as a collection.

As in the examples in the previous sections, we now can use a Data Table to execute simulation trials and gather sample statistics. Figure 13.32 shows the results of 1,000 simulation trials. PTW can use this simulation model to assess its exposure to currency exchange rates and consider actions to hedge against this risk. For example, PTW can assess the probability of total end-of-quarter revenue in U.S. dollars being below a certain threshold. Figure 13.32 shows that there is approximately a 0.287 probability that total end-of-quarter revenue in U.S. dollars will be below $475,000 (with a 95% confidence interval of 0.259 to 0.315).

MODEL file
quarterlyexchange

Figure 13.32 Output from PTW Simulation

	A	B	C	D	E
1	**Press Teag Worldwide**				
2					
3	**Parameters**				
4	Start-of-Quarter Exchange Rate (per $)	£0.615	NZD 1.200	¥87.10	Random Scenario
5	Quarterly % Change in Exchange Rate	2.08%	3.64%	9.11%	252
6	End-of-Quarter Exchange Rate (per $)	£0.628	NZD 1.244	¥95.03	
7	Quarterly Revenue	£100,000	NZD 250,000	¥10,000,000	
8					
9	**Model**				Total
10	End-of-Quarter Revenue in $	$159,237	$201,016	$105,225	$465,478
11				Total Revenue ($1000):	$465
12					
13	Simulation Trial	Total Revenue ($1000)		Total Revenue ($1000) Summary Statistics	
14	1	$465		Count	1000
15	2	$517		Minimum	$390
16				Maximum	$586
17				Mean	$486
18				Standard Deviation	$22
19					
20				95% C.I. on Mean Revenue	$484
21					$487
22					
23				95% P.I. on Revenue	$442
24					$530
25					
26				P(Revenue < $475)	0.287
27				Standard Error of Proportion	0.014
28					
29				95% C.I. on P(Revenue < $475)	0.259
30					0.315



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notes + Comments

In this section, to capture dependency between random variables, we directly sampled the same scenario to generate a value of each random variable. This approach has the benefit of capturing all types of dependency (not just linear relationships between random variables).

In general, it is also possible to create dependency between random variables from (possibly different) fitted distributions. However, doing this in native Excel is beyond the scope of this book. Specialized simulation software such as @RISK, Crystal Ball, and Analytic Solver, provide automated procedures to incorporate dependency between random variables.

13.5 Simulation Considerations

Verification and Validation

An important aspect of any simulation study involves confirming that the simulation model accurately describes the real system. Inaccurate simulation models cannot be expected to provide worthwhile information. Thus, before using simulation results to draw conclusions about a real system, one must take steps to verify and validate the simulation model.

Verification is the process of determining that the computer procedure that performs the simulation calculations is logically correct. Verification is largely a debugging task to make sure that there are no errors in the computer procedure that implements the simulation. In some cases, an analyst may compare computer results for a limited number of events with independent hand calculations. In other cases, tests may be performed to verify that the random variables are being generated correctly and that the output from the simulation model seems reasonable. The verification step is not complete until the user develops a high degree of confidence that the computer procedure is error free.

Validation is the process of ensuring that the simulation model provides an accurate representation of a real system. Validation requires an agreement among analysts and managers that the logic and the assumptions used in the design of the simulation model accurately reflect how the real system operates. The first phase of the validation process is done prior to, or in conjunction with, the development of the computer procedure for the simulation process. Validation continues after the computer program has been developed, with the analyst reviewing the simulation output to see whether the simulation results closely approximate the performance of the real system. If possible, the output of the simulation model is compared to the output of an existing real system to make sure that the simulation output closely approximates the performance of the real system. If this form of validation is not possible, an analyst can experiment with the simulation model and have one or more individuals experienced with the operation of the real system review the simulation output to determine whether it is a reasonable approximation of what would be obtained with the real system under similar conditions.

Verification and validation are not tasks to be taken lightly. They are key steps in any simulation study and are necessary to ensure that decisions and conclusions based on the simulation results are appropriate for the real system.

Advantages and Disadvantages of Using Simulation

The primary advantages of simulation are that it is conceptually easy to understand and that the methods can be used to model and learn about the behavior of complex systems that would be difficult, if not impossible, to deal with analytically. Simulation models are flexible; they can be used to describe systems without requiring the assumptions that are often required by other mathematical models. In general, the larger the number of random variables a system has, the more likely it is that a simulation model will provide the best approach for studying the system. Another advantage is that a simulation model provides a convenient experimental laboratory for the real system. Changing assumptions or operating policies in

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the simulation model and rerunning it can provide results that help predict how such changes will affect the operation of the real system. Experimenting directly with a real system is often not feasible. Simulation models frequently warn against poor decision strategies by projecting disastrous outcomes such as system failures, large financial losses, and so on.

Simulation is not without disadvantages. For complex systems, the process of developing, verifying, and validating a simulation model can be time consuming and expensive. However, the process of developing the model generally leads to a better understanding of the system, which is an important benefit. Like all mathematical models, the analyst must be conscious of the assumptions of the model in order to understand its limitations. In addition, each simulation run provides only a sample of output data. As such, the summary of the simulation data provides only estimates or approximations about the real system. Nonetheless, the danger of obtaining poor solutions is greatly mitigated if the analyst exercises good judgment in developing the simulation model and follows proper verification and validation steps. Furthermore, if a sufficiently large enough set of simulation trials is run under a wide variety of conditions, the analyst will likely have sufficient data to predict how the real system will operate.

Summary

Simulation is a method for learning about a real system by experimenting with a model that represents the system. Some of the reasons simulation is frequently used are as follows:

1. It can be used for a wide variety of practical problems.
2. The simulation approach is relatively easy to explain and understand. As a result, management confidence is increased and the results are more easily accepted.
3. Spreadsheet software such as Excel and specialized software packages have made it easier to develop and implement simulation models for increasingly complex problems.

In this chapter, we showed how native Excel functions can be used to execute simulation models on several examples. For the Sanotronics problem, we used simulation to evaluate the risk involving the development of a new product. For the Promus Corp problem, we used a simulation model to evaluate different inventory policies for the company to use when demand is uncertain. Then we developed a simulation model to help Land Shark Inc. estimate how varying its bid amount affects the likelihood of winning a property auction. We then demonstrated how to build a simulation model with dependent random variables using the Press Teag Worldwide example that included correlated fluctuations for currency exchange rates. With the steps below, we summarize the procedure for developing a simulation model involving controllable inputs, uncertain inputs represented by random variables, and output measures.

Summary of Steps for Conducting a Simulation Analysis

Excel's Auditing Tools are introduced in Chapter 12.

1. *Construct a spreadsheet model that computes output measures for given values of inputs.* The foundation of a good simulation model is logic that correctly relates input values to outputs. Audit the spreadsheet to ensure that the cell formulas correctly evaluate the outputs over the entire range of possible input values.
2. *Identify inputs that are uncertain, and specify probability distributions for these cells* (rather than just static numbers). Note that all inputs may not have a degree of uncertainty sufficient to require modeling with a probability distribution. Other inputs may actually be decision variables, which are not random and should not be modeled with probability distributions; rather, these are values that the decision maker can control.
3. *Select one or more outputs to record over the simulation trials.* Typical information recorded for an output includes a histogram of output values over all simulation trials and summary statistics such as the mean, standard deviation, maximum, minimum, and percentile values.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

4. *Execute the simulation for a specified number of trials.* In this chapter, we have used 1,000 trials for our simulation models. The amount of sampling error can be monitored by observing how much simulation output measures fluctuate across multiple simulation runs. If the confidence intervals on the output measures are unacceptably wide, the number of trials can be increased to reduce the amount of sampling error.
5. *Analyze the outputs and interpret the implications on the decision-making process.* In addition to estimates of the mean output, simulation allows us to construct a distribution of possible output values. Analyzing the simulation results allows the decision maker to draw conclusions about the operation of the real system.

In this chapter, we have focused on Monte Carlo simulation consisting of independent trials in which the results for one trial do not affect what happens in subsequent trials. Another style of simulation, called **discrete-event simulation**, involves trials that represent how a system evolves over time. One common application of discrete-event simulation is the analysis of waiting lines. In a waiting-line simulation, the random variables are the interarrival times of the customers and the service times of the servers, which together determine the waiting and completion times for the customers. Although it is possible to conduct small discrete-event simulations with native Excel functionality, discrete-event simulation modeling is best conducted with special-purpose software such as Arena®, ProModel®, and Simio®. These packages have built-in simulation clocks, simplified methods for generating random variables, and procedures for collecting and summarizing the simulation output.

Glossary

Base-case scenario Output resulting from the most likely values for the random variables of a model.
Best-case scenario Output resulting from the best values that can be expected for the random variables of a model.
Continuous probability distribution A probability distribution for which the possible values for a random variable can take any value in an interval or collection of intervals. An interval can include negative and positive infinity.
Controllable input Input to a simulation model that is selected by the decision maker.
Discrete probability distribution A probability distribution for which the possible values for a random variable can take on only specified discrete values.
Discrete-event simulation A simulation method that describes how a system evolves over time by using events that occur at discrete points in time.
Monte Carlo simulation A simulation method that uses repeated random sampling to represent uncertainty in a model representing a real system and that computes the values of model outputs.
Probability distribution A description of the range and relative likelihood of possible values of a random variable (uncertain quantity).
Random variable (uncertain variable) Input to a simulation model whose value is uncertain and described by a probability distribution.
Risk analysis The process of evaluating a decision in the face of uncertainty by quantifying the likelihood and magnitude of an undesirable outcome.
Trial A set of values for the random variables and the associated values of output measures for a single scenario in a simulation model.
Validation The process of determining that a simulation model provides an accurate representation of a real system.
Verification The process of determining that a computer program implements a simulation model as it is intended.
What-if analysis A trial-and-error approach to learning about the range of possible outputs for a model. Trial values are chosen for the model inputs (these are the what-ifs) and the value of the output(s) is computed.
Worst-case scenario Output resulting from the worst values that can be expected for the random variables of a model.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Problems

MODEL file
galaxy

1. **Virtual Reality Goggle Inventory.** Galaxy Co. sells virtual reality (VR) goggles targeted to customers who like to play video games. Galaxy procures each pair of goggles for $150 from its supplier and sells each pair of goggles for $300. Monthly demand for the VR goggles is a normal random variable with a mean of 160 units and a standard deviation of 40 units. At the beginning of each month, Galaxy orders enough goggles from its supplier to bring the inventory level up to 140 goggles. If the monthly demand is less than 140, Galaxy pays $20 per pair of goggles that remains in inventory at the end of the month. If the monthly demand exceeds 140, Galaxy sells only the 140 pairs of goggles in stock. Galaxy assigns a shortage cost of $40 for each unit of demand that is unsatisfied to represent a loss-of-goodwill among its customers. **LO 2**
   a. Construct a simulation model to compute the monthly profit and unmet demand for Galaxy Co.
   b. What is the average monthly profit resulting from its policy of stocking 140 pairs of goggles at the beginning of each month?
   c. What is the proportion of months in which demand is completely satisfied?
   d. Compare the average monthly profit resulting from stocking 105, 140, and 175 pairs of goggles at the beginning of each month. Which amount results in the largest average monthly profit?

MODEL file
dice

2. **Dice Rolls.** Construct a spreadsheet simulation model to simulate 1,000 rolls of a six-sided die with the six sides numbered 1, 2, 3, 4, 5, and 6. **LO 3**
   a. Construct a histogram of the 1,000 observed dice rolls by counting the number of times the roll is 1, 2, 3, 4, 5, and 6.
   b. For each roll of two dice, record the sum of the dice. Construct a histogram of the 1,000 observations of the sum of two dice by counting the number of times the roll is 2, 3, 4, . . ., 12.
   c. For each roll of three dice, record the sum of the dice. Construct a histogram of the 1,000 observations of the sum of three dice by counting the number of times the roll is 3, 4, 5, . . ., 18.
   d. For each roll of four dice, record the sum of the dice. Construct a histogram of the 1,000 observations of the sum of four dice by counting the number of times the roll is 4, 5, 6, . . ., 24.
   e. Compare the histograms in parts (a), (b), (c), and (d). What statistical phenomenon does this sequence of charts illustrate?

MODEL file
madeira

3. **Wearable Electronic Product Launch.** The management of Madeira Computing is considering the introduction of a wearable electronic device with the functionality of a laptop computer and phone. The fixed cost to launch this new product is $300,000. The variable cost for the product is expected to be between $160 and $240, with a most likely value of $200 per unit. The product will sell for $300 per unit. Demand for the product is expected to range from 0 to approximately 20,000 units, with 4,000 units the most likely. **LO 1, 2**
   a. Compute profit for this product in the base-case scenario.
   b. Compute profit for this product in the worst-case scenario.
   c. Compute profit for this product in the best-case scenario.
   d. Model the variable cost as a uniform random variable with a minimum of $160 and a maximum of $240. Model the product demand as 1,000 times the value of a gamma random variable with an alpha parameter of 3 and a beta parameter of 2. Construct a simulation model to estimate the average profit and the probability that the project will result in a loss.
   e. What is your recommendation regarding whether to launch the product?

MODEL file
brinkley

4. **Profitability of New Product.** The management of Brinkley Corporation is interested in using simulation to estimate the profit per unit for a new product. The selling price

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

for the product will be $45 per unit. Probability distributions for the purchase cost, the labor cost, and the transportation cost are estimated as follows. LO 1, 3

Procurement Cost ($)	Probability	Labor Cost ($)	Probability	Transportation Cost ($)	Probability
10	0.25	20	0.10	3	0.75
11	0.45	22	0.25	5	0.25
12	0.30	24	0.35		
		25	0.30		

a. Compute profit per unit for the base-case scenario.
b. Compute profit per unit for the worst-case scenario.
c. Compute profit per unit for the best-case scenario.
d. Construct a simulation model to estimate the average profit per unit. What is a 95% confidence interval around this average?
e. Why is the simulation approach to risk analysis preferable to generating a variety of what-if scenarios?
f. Management believes that the project may not be sustainable if the profit per unit is less than $5. Use the simulation model from part (d) to estimate the probability that the profit per unit will be less than $5. What is a 95% confidence interval around this proportion?

MODEL file
statewide

5. **Estimating Auto Accident Costs.** Statewide Auto Insurance believes that for every trip longer than 10 minutes that a teenager drives, there is a 1 in 1,000 chance that the drive will result in an auto accident. Assume that the cost of an accident can be modeled with a beta distribution with an alpha parameter of 1.5, a beta parameter of 3, a minimum value of $500, and a maximum value of $20,000. Statewide is interested in analyzing the total cost resulting from accidents that may occur when a teenager drives 500 trips longer than 10 minutes. LO 2, 3
   a. In addition to the cost of each accident being uncertain, the number of accidents that occur when a teenager drives 500 trips longer than 10 minutes is also uncertain. What type of random variable is appropriate for modeling the number of accidents in 500 trips?
   b. Construct a simulation model incorporating the uncertainty in the number of accidents in 500 trips and the uncertain cost of each accident.
   c. If a teenager drives 500 trips longer than 10 minutes, what is the average cost resulting from accidents? Provide a 95% confidence interval on this mean.
   d. If a teenager drives 500 trips longer than 10 minutes, what is the probability that the total cost from accidents will exceed $8,000? Provide a 95% confidence interval on this proportion.

MODEL file
statefarm

6. **Automobile Collision Claims.** State Farm Insurance has developed the following table to describe the distribution of automobile collision claims paid during the past year. LO 3
   a. Set up a table of number intervals that can be used with the Excel XLOOKUP function and random numbers to generate values for automobile collision claim payments.
   b. Construct a simulation model to estimate the average claim payment amount and the standard deviation in the claim payment amounts.
   c. Let X be the discrete random variable representing the dollar value of an automobile collision claim payment. Let, $x_1, x_2, \ldots, x_n$ represent possible values of X. Then, the mean (μ) and standard deviation (σ) of X can be computed as $\mu = x_1 \times P(X = x_1) + \cdots + x_n \times P(X = x_n)$, and $\sigma = \sqrt{(x_1 - \mu)^2 \times P(X = x_1) + \cdots + (x_n - \mu)^2 \times P(X = x_n)}$. Compare the values of sample mean and sample standard deviation in part (b) to the analytical

In Chapter 5, we describe the analytical calculation of the mean and standard deviation of a random variable.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

calculation of the mean and standard deviation. How can we improve the accuracy of the sample estimates from the simulation?

Payment ($)	Probability
0	0.83
500	0.06
1,000	0.05
2,000	0.02
5,000	0.02
8,000	0.01
10,000	0.01

MODEL file
playoffs

7. **Playoff Series in National Basketball Association.** The Dallas Mavericks and the Golden State Warriors are two teams in the National Basketball Association (NBA). Dallas and Golden State will play multiple times over the course of an NBA season. Assume that the Dallas Mavericks have a 25% probability of winning each game against the Golden State Warriors. LO 3
   a. Construct a simulation model that uses the negative binomial distribution to simulate the number of games Dallas would lose before winning four games against the Golden State Warriors.
   b. Now suppose that the Dallas Mavericks face the Golden State Warriors in a best-of-seven playoff series in which the first team to win four games out of seven wins the series. Using the simulation model from part (a), estimate that probability that the Dallas Mavericks would win a best-of-seven series against the Golden State Warriors.

MODEL file
grear

8. **Tire Warranty Analysis.** Grear Tire Company has produced a new tire with an estimated mean lifetime mileage of 36,500 miles. Management also believes that the standard deviation is 5,000 miles and that tire mileage is normally distributed. To promote the new tire, Grear has offered to refund some money if the tire fails to reach 30,000 miles before the tire needs to be replaced. Specifically, for tires with a lifetime below 30,000 miles, Grear will refund a customer $1 per 100 miles short of 30,000. Construct a simulation model to answer the following questions. LO 2
   a. For each tire sold, what is the average cost of the promotion?
   b. What is the probability that Grear will refund more than $25 for a tire?

MODEL file
gustin

9. **Yield from Recruiting Seminars.** To generate leads for new business, Gustin Investment Services offers free financial planning seminars at major hotels in Southwest Florida. Gustin conducts seminars for groups of 25 individuals. Each seminar costs Gustin $3,500, and the commission for each new account opened is $5,000. Gustin estimates that for each individual attending the seminar, there is a 0.01 probability that they will open a new account. LO 3
   a. For given values of the relevant input parameters, determine the equation for computing Gustin's profit per seminar.
   b. What type of random variable is the number of new accounts opened? *Hint*: Review Appendix 13.1 for descriptions of various types of probability distributions.
   c. Construct a simulation model to analyze the profitability of Gustin's seminars. Would you recommend that Gustin continue running the seminars?
   d. Using the simulation model in part (c), what is the minimum number of attendees (in a multiple of five, i.e., 25, 30, 35, etc.) that Gustin needs so that the seminar's average profit is greater than zero?

MODEL file
landsharkbeta

10. **Analysis of Bid Amounts.** Using the file *landsharkbeta*, evaluate bid amounts from $1,239,000 to $1,329,000 in increments of $30,000 by building a table listing 95% confidence intervals around the average return and probability of winning the auction. Which of these bid amounts do you recommend? LO 2, 3

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

MODEL file
gleague

11. **Point Distribution in a Basketball Game.** The Iowa Wolves are scheduled to play against the Maine Red Claws in an upcoming game in the National Basketball Association (NBA) G League. Because a player in the NBA G League is still developing their skills, the number of points a player scores in a game can vary substantially. Assume that each player's point production can be represented as an integer uniform random variable with the ranges provided in the following table. LO 3

Player	Iowa Wolves	Maine Red Claws
1	[5,20]	[7,12]
2	[7,20]	[15,20]
3	[5,10]	[10,20]
4	[10,40]	[15,30]
5	[6,20]	[5,10]
6	[3,10]	[1,20]
7	[2,5]	[1,4]
8	[2,4]	[2,4]

a. Develop a spreadsheet model that simulates the points scored by each team and the difference in their point totals.
b. What are the average and standard deviation of points scored by the Iowa Wolves? What is the shape of the distribution of points scored by the Iowa Wolves?
c. What are the average and standard deviation of points scored by the Maine Red Claws? What is the shape of the distribution of points scored by the Maine Red Claws?
d. Let Point Differential = Iowa Wolves points − Maine Red Claws points. What is the average Point Differential between the Iowa Wolves and Maine Red Claws? What is the standard deviation of the Point Differential? What is the shape of the Point Differential distribution?
e. What is the probability that the Iowa Wolves score more points than the Maine Red Claws?
f. The coach of the Iowa Wolves feels that they are the underdog and is considering a riskier game strategy. The effect of this strategy is that the range of each Wolves player's point production increases symmetrically so that the new range is [0, original upper bound + original lower bound]. For example, Wolves player 1's range with the risky strategy is [0, 25]. How does the new strategy affect the average and standard deviation of the Wolves point total? How does that affect the probability that the Iowa Wolves score more points than the Maine Red Claws?

MODEL file
nyse

12. **Simulating Stock Price.** Suppose that the price of a share of a particular stock listed on the New York Stock Exchange (NYSE) is currently $39. The following probability distribution shows how the price per share is expected to change over a three-month period. LO 3

Stock Price Change ($)	Probability
−2	0.05
−1	0.10
0	0.25
+1	0.20
+2	0.20
+3	0.10
+4	0.10

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Construct a spreadsheet simulation model that computes the value of the stock price in 3 months, 6 months, 9 months, and 12 months under the assumption that the change in stock price over any three-month period is independent of the change in stock price over any other three-month period. For a current price of $39 per share, what is the average stock price per share 12 months from now? What is the standard deviation of the stock price 12 months from now?
b. Based on the model assumptions, what are the lowest and highest possible prices for this stock in 12 months? Based on your knowledge of the stock market, how valid do you think this is? Propose an alternative to modeling how stock prices evolve over three-month periods.

MODEL file
overbooking

13. **Airline Overbooking.** Allegiant Airlines is considering an overbooking policy for one of its flights. The airplane has 50 seats, but Allegiant is considering accepting more reservations than seats because sometimes passengers do not show up for their flights, resulting in empty seats. The *Data* worksheet in the file *overbooking* contains data on 1,000 passengers showing whether they showed up for their respective flights.

In addition, Allegiant has conducted a field experiment to gauge the demand for reservations for the current flight. During this experiment, they did not limit the number of reservations for the flight to observe the uncensored demand. The following table summarizes the result of the field experiment.

Number of Reservations Demanded	Probability
48	0.05
49	0.05
50	0.15
51	0.30
52	0.25
53	0.10
54	0.10

Allegiant receives a marginal profit of $100 for each passenger who books a reservation (regardless of whether they show up). Allegiant incurs a rebooking cost of $300 for each passenger who books a reservation, but is denied seating due to a full airplane; this cost results from rescheduling the passenger and any loss of goodwill. LO 3

a. Model the number of booked passengers who show up for their flight as a binomial random variable. Describe how to determine the two parameters (n and p) for the binomial distribution in this application. *Hint:* Review Appendix 13.1 for description of the binomial distribution and consider the the data in the *Data* worksheet of the file *overbooking*.
b. The number of reservations demanded is a random variable. How should it be modeled?
c. To control its rebooking costs, Allegiant wants to set a limit on the number of reservations it will accept. Evaluate Allegiant's average net profit for reservation limits of 50, 52, and 54, respectively. Based on the 95% confidence intervals for average net profit, which reservation limit do you recommend?

14. **Project Management.** A project has four activities (A, B, C, and D) that must be performed sequentially. The probability distributions for the time required to complete each of the activities are as follows. Construct a simulation model to answer the following questions. LO 3

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

MODEL file
project

Activity	Activity Time (Weeks)	Probability
A	5	0.25
	6	0.35
	7	0.25
	8	0.15
B	3	0.20
	5	0.55
	7	0.25
C	10	0.10
	12	0.25
	14	0.40
	16	0.20
	18	0.05
D	8	0.60
	10	0.40

a. What is the average length of the project and the standard deviation of the project length.
b. Using the simulation model in part (a), what is the estimated probability that the project will be completed in 35 weeks or less?

MODEL file
ftc

15. **Holiday Toy Inventory.** In preparing for the upcoming holiday season, Fresh Toy Company (FTC) designed a new doll called The Dougie that teaches children how to dance. The fixed cost to produce the doll is $100,000. The variable cost, which includes material, labor, and shipping costs, is $34 per doll. During the holiday selling season, FTC will sell the dolls for $42 each. If FTC overproduces the dolls, the excess dolls will be sold in January through a distributor who has agreed to pay FTC $10 per doll. Demand for new toys during the holiday selling season is uncertain. The normal probability distribution with an average of 60,000 dolls and a standard deviation of 15,000 is assumed to be a good description of the demand. FTC has tentatively decided to produce 60,000 units (the same as average demand), but it wants to conduct an analysis regarding this production quantity before finalizing the decision. **LO 2**
a. Create a what-if spreadsheet model using formulas that relate the values of production quantity, demand, sales, revenue from sales, amount of surplus, revenue from sales of surplus, total cost, and net profit. What is the profit when demand is equal to its average (60,000 units)?
b. Modeling demand as a normal random variable with a mean of 60,000 and a standard deviation of 15,000, simulate the sales of The Dougie doll using a production quantity of 60,000 units. What is the estimate of the average profit associated with the production quantity of 60,000 dolls? How does this compare to the profit corresponding to the average demand as computed in part (a)?
c. Before making a final decision on the production quantity, management wants an analysis of a more aggressive 70,000-unit production quantity and a more conservative 50,000-unit production quantity. Run your simulation with these two production quantities. What is the average profit associated with each?
d. Besides average profit, what other factors should FTC consider in determining a production quantity? Compare the four production quantities (40,000; 50,000; 60,000; and 70,000) using all these factors. What trade-offs occur? What is your recommendation?

MODEL file
contractor

16. **Contractor Bidding.** Jonah Arkfeld, a building contractor, is preparing a bid on a new construction project. Two other contractors will be submitting bids for the same project. Jonah has analyzed past bidding practices and the requirements of the project to determine

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the probability distributions of the two competing contractors. The bid from Contractor A can be described with a triangular distribution with a minimum value of $600,000, a maximum value of $800,000, and a most likely value of $725,000. The bid from Contractor B can be described with a normal distribution with a mean of $700,000 and a standard deviation of $50,000. Construct a simulation model to answer the following questions. **LO 2**

a. If Jonah submits a bid of $750,000, what is the probability that Jonah will have the lowest bid and win the project?
b. What is the probability that Contractor A has the lowest bid and wins the project?
c. What is the probability that Contractor B has the lowest bid and wins the contract?

MODEL file
hybrid

17. **Hybrid Car Analysis.** You are considering the purchase of a new car and are weighing the choice between a hybrid sedan (which assists a gasoline engine with an electric motor powered via regenerative braking) and the standard gasoline-powered (non-hybrid) sedan. The non-hybrid version costs $23,240 with fuel economy of 21 miles per gallon in city driving and 32 miles per gallon in highway driving. The hybrid version of the car costs $25,990 with fuel economy of 43 miles per gallon in city driving and 41 miles per gallon in highway driving.

You plan to keep the car for 10 years. Your annual mileage is uncertain; you only know that each year you will drive between 9,000 and 13,000 miles. Based on your past driving patterns, 60% of your miles are city driving and 40% of your miles are highway driving. The current gasoline price is $2.19 per gallon, but you know that gasoline prices vary unpredictably over time.

Compute the net present value (NPV) of the costs of each vehicle (purchase cost + gasoline cost) using a discount rate of 3%. Assume that you pay the entire purchase price of the vehicle immediately (Year 0) and the annual gasoline costs are incurred at the end of each year. Construct a simulation model to answer the following questions. **LO 2, 4, 5**

a. Does the gas price in year t+1 depend on the gas price in year t? Does the percent change in gas price from year t to year t+1 depend on the percent change in gas price from year t-1 to year t? What do these observations suggest about the way to model gas prices over a period of time in a simulation model? *Hint:* Use the Excel formula =CORREL(C2:C87, C3:C88) to compute the correlation between the gas prices in consecutive years. A similar approach can be taken to calculate the correlation of percent change in gas price in consecutive years.
b. Construct a simulation model to determine, on average, what is the average cost savings of the hybrid vehicle over the non-hybrid?
c. Because of your concern about the maintenance needs of the hybrid vehicle, you would need to be assured of significant savings to convince you to purchase the hybrid. Using the simulation model in part (b), what is the probability that the hybrid will result in more than $2,000 in savings over the non-hybrid?

MODEL file
orange

18. **Product Adoption.** Orange Tech (OT) is a software company that provides a suite of programs that are essential to everyday business computing. OT has just enhanced its software and released a new version of its programs. For financial planning purposes, OT needs to forecast its revenue over the next few years. To begin this analysis, OT is considering one of its largest customers. Over the planning horizon, assume that this customer will upgrade at most once to the newest software version, but the number of years that pass before the customer purchases an upgrade varies. Up to the year that the customer actually upgrades, assume there is a 0.50 probability that the customer upgrades in any particular year. In other words, the upgrade year of the customer is a (discrete) random variable. For guidance on an appropriate way to model upgrade year, refer to Appendix 13.1. Furthermore, the revenue that OT earns from the customer's upgrade also varies (depending on the number of programs the customer decides to upgrade). Assume that the revenue from an upgrade obeys a normal distribution with a mean of $100,000 and a standard deviation of $25,000. **LO 2, 3**

a. Construct a simulation model that computes the net present value of the revenue from the customer upgrade. *Hint:* Excel's NPV function computes the net present value for

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a sequence of cash flows that occur at the end of each period. To correctly use this function for cash flows that occur at the beginning of each period, use the formula =NPV(*discount rate, flow range*) + *initial amount*, where *discount rate* is the annual discount rate, *flow range* is the cell range containing cash flows for years 1 through *n*, and *initial amount* is the cash flow in the initial period (year 0). Use an annual discount rate of 10%.

b. What is the average net present value that OT earns from this customer?

c. What is the standard deviation of net present value? How does this compare to the standard deviation of the revenue? Explain.

MODEL file vaccine

19. **Flu Vaccine Ordering.** OuRx, a retail pharmacy chain, is faced with the decision of how much flu vaccine to order for the next flu season. OuRx has to place a single order for the flu vaccine several months before the beginning of the season because it takes four to five months for the supplier to create the vaccine. OuRx wants to more closely examine the ordering decision because, over the past few years, the company has ordered too much vaccine or too little. OuRx pays a wholesale price of $12 per dose to obtain the flu vaccine from the supplier and then sells the flu shot to their customers at a retail price of $20.

Because OuRx earns a profit on flu shots that it sells and it can't sell more than its supply, the appropriate profit computation depends on whether demand exceeds the order quantity or vice versa. Similarly, the number of lost sales and excess doses depends on whether demand exceeds the order quantity or vice versa. Demand for the flu vaccine is uncertain. The *VaccineDemand* worksheet in the file *vaccine* contains data produced by epidemiologists to help OuRx gain insight on demand for flu vaccine at their retail pharmacies. **LO 2**

a. Construct a base spreadsheet model that correctly computes net profit for a given level of demand and specified order quantity. For an order quantity of 500,000 doses, what is the net profit when demand is 400,000 doses and 600,000 doses, respectively?

b. To help determine how to model flu vaccine demand, construct a histogram of the data provided in the *VaccineDemand* worksheet in the file *vaccine*. In column B, compute the natural logarithm (using the Excel function LN) of each observation and construct a histogram of these logged demand observations. Based on the histograms of the non-logged demand and logged demand, respectively, what seems to be a good choice of probability distribution for (non-logged) vaccine demand? *Hint:* Review Appendix 13.1 for descriptions of various types of probability distributions.

c. Representing flu vaccine demand with the type of random variable you identified in part (b), complete the simulation model and determine the average net profit resulting from an order quantity of 500,000 doses. What is the 95% confidence interval on the average profit? What is the probability of running out of the flu vaccine?

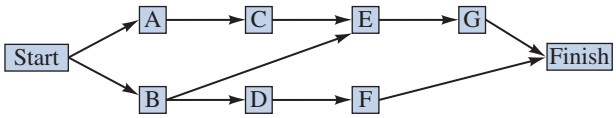
MODEL file scope

20. **Music Concert Event Management.** At a local university, the Student Commission on Programming and Entertainment (SCOPE) is preparing to host its first music concert of the school year. To successfully produce this music concert, SCOPE has to complete several activities. The following table lists information regarding each activity. An activity's immediate predecessors are the activities that must be completed before the considered activity can begin. The table also lists duration estimates (in days) for each activity.

Activity	Immediate Predecessors	Minimum Time	Likely Time	Maximum Time
A: Negotiate contract with selected musicians	—	5	6	9
B: Reserve site	—	8	12	15
C: Logistical arrangements for music group	A	5	6	7
D: Screen and hire security personnel	B	3	3	3
E: Advertising and ticketing	B, C	1	5	9
F: Hire parking staff	D	4	7	10
G: Arrange concession sales	E	3	8	10

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The following network illustrates the precedence relationships in the SCOPE project. The project begins with activities A and B, which can start immediately (time 0) because they have no predecessors. On the other hand, activity E cannot be started until activities B and C are both completed. The project is not complete until all activities are completed. **LO 2**



a. Using the triangular distribution to represent the duration of each activity, construct a simulation model to estimate the average amount of time to complete the concert preparations.
b. What is the likelihood that the project will be complete in 23 days or less?

MODEL file
jets

21. **Airplane Maintenance.** Steve Austin is the fleet manager for SharePlane, a company that sells fractional ownership of private jets. SharePlane must carefully maintain their jets at all times. If a jet breaks down, it must be repaired immediately. Even if a jet functions well, it must be maintained at regularly scheduled intervals. Currently, Steve is managing two jets, Jet A and Jet B, for a collection of clients and is interested in estimating their availability in between trips to the repair shop as having both jets out-of-service due to repair or maintenance at the same time can affect its customer service. Jet A and Jet B have just completed preventive maintenance. The next maintenance is scheduled for both Jet A and Jet B in four months. It is also possible that one or both will break down before this scheduled maintenance and require repair. The amount of time to a plane's first failure is uncertain. Historical data recording the time to a plane's first failure (measured in months) is provided in the *TimeToFailData* worksheet of the file *jets*. Determine an appropriate probability distribution for these data. Furthermore, once a plane enters repair (either due to a failure or as scheduled maintenance), the amount of time the plane will be in maintenance is also uncertain. Historical data recording the repair time (measured in months) is provided in the *RepairTimeData* worksheet of the file *jets*. Examine the appropriateness of fitting a log-normal distribution to these data. Steve wants to develop a simulation model to estimate the length of time that Jet A and Jet B are both out-of-service over the next few months. For simplicity, you can assume that these planes will enter repair or maintenance just once over the next few months. **LO 2**

a. Construct a simulation model that computes the length of time that Jet A and Jet B are both out-of service over the next few months.
b. What is the average amount of time that the planes are both out-of-service?
c. What is the probability that the planes are both out-of-service for longer than 1.5 months?

MODEL file
blackjack

22. **Blackjack Card Game.** Blackjack, or 21, is a popular casino game that begins with each player and the dealer being dealt two cards. The value of each hand is determined by the point total of the cards in the hand. Face cards and 10s count 10 points, aces can be counted as either 1 or 11 points, and all other cards count at their face value. For instance, the value of a hand consisting of a jack and an 8 is 18; the value of a hand consisting of an ace and a two is either 3 or 13, depending on whether player counts the ace as 1 or 11 points. The goal is to obtain a hand with a value as close as possible to 21 without exceeding 21. After the initial deal, each player and the dealer may draw additional cards (called "taking a hit") in order to improve their hand. If a player or the dealer takes a hit and the value of the hand exceeds 21, that person "goes broke" and loses. The dealer's advantage is that each player must decide whether to take a hit before the dealer decides whether to take a hit. If a player takes a hit and goes over 21, the player loses even if the dealer later takes a hit and goes over 21. For

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

this reason, players will often decide not to take a hit when the value of their hand is 12 or greater.

The dealer's hand is dealt with one card up (face showing) and one card down (face hidden). The player then decides whether to take a hit based on knowledge of the dealer's up card. Suppose that you are playing blackjack and the dealer's up card is a 6 and your hand has a value of 16 for the two cards initially dealt.

With a hand of a value of 16, if you decide to take a hit, the following cards will improve your hand: ace, 2, 3, 4, or 5. Any card with a point count greater than 5 will result in you going broke. Assume that if you have a hand with a value of 16, the following probabilities describe the ending value of your hand:

Value of Hand	17	18	19	20	21	Broke
Probability	0.0769	0.0769	0.0769	0.0769	0.0769	0.6155

A gambling professional determined that when the dealer's up card is a 6, the following probabilities describe the ending value of the dealer's hand:

Value of Hand	17	18	19	20	21	Broke
Probability	0.1654	0.1063	0.1063	0.1017	0.0972	0.4231

Construct a simulation model to simulate the result of 1,000 blackjack hands when the dealer has a 6 up and you take a hit with a hand that has a value of 16. **LO 3**

a. What is the probability of the dealer winning, a push (a tie), and you winning, respectively?
b. If you have a hand with a value of 16 and don't take a hit, the only way that you can win is if the dealer goes broke. If you don't take a hit, what is the probability of the dealer winning, a push (a tie), and you winning, respectively?
c. Based on the results from parts (a) and (b), should you take a hit or not if you have a hand of value 16 and the dealer has a 6 up?

MODEL file
birthday

23. **Birthday Paradox.** As an ice-breaker on the first day of analytics class with 25 students in attendance, a professor announces that a free textbook will be awarded to any students who share a birthday (month and day, not necessarily same year). Assume the probability of any student being born on any particular day is 1 in 365 (neglecting leap years). **LO 3**
a. Construct a simulation model to estimate the probability that there is at least one pair of students sharing the same birthday.
b. What is a 95% confidence interval on the probability provided in part (a)? .

MODEL file
jackpot

24. **Distribution of Money in Jackpot Soap Bars.** A creative entrepreneur has created a novelty soap called Jackpot. Inside each bar of Jackpot soap is a rolled-up bill of U.S. currency. There are 1,000 bars of soap in the initial offering of the soap. Although the denomination of the bill inside a bar of soap is unknown, the distribution of bills in these first 1,000 bars is given in the following table:

Bill Denomination ($)	Number of Bills
1	520
5	260
10	130
20	60
50	29
100	1
Total	1,000

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If a customer buys 40 bars of soap, the number of bars that contain a $50 or $100 bill is uncertain. LO 3

a. What type of random variable describes the number of bars (out of 40 bars) that contain a $50 or $100 bill? *Hint:* Review Appendix 13.1 for descriptions of various types of probability distributions.
b. Construct a model to simulate the number of bars that contain a $50 or $100 bill.
c. On average, how many of these bars contain a $50 or $100 bill?
d. What is the probability that at least one of the 40 bars contains a $50 or $100 bill?

MODEL file
jackpot2

25. **Jackpot Soap (Revisited).** Refer to the Jackpot soap scenario in Problem 24. After the sale of the original 1,000 bars of soap, Jackpot soap went viral, and the soap has become wildly popular. Production of the soap has been ramped up so that now millions of bars have been produced. However, the distribution of the bills in the soap obeys the same distribution as outlined in Problem 24. A customer is interested in analyzing the number of bars that they need to buy so that the last bar purchased is the third bar to contain a bill of at least $20 value. This required number of bars is uncertain. LO 3

a. What type of random variable describes the number of bars containing a bill of less than $20 value that are purchased before obtaining the third bar to contain a bill of at least $20 value? *Hint:* Review Appendix 13.1 for descriptions of various types of probability distributions.
b. Construct a model to simulate the number of bars that need to be purchased so that the last bar purchased is the third bar to contain a bill of at least $20 value.
c. On average, how many bars of soap will a customer need to buy before obtaining three bars of soap containing a bill of at least $20 value?
d. What is the probability that the customer will need to buy more 50 bars of soap?

MODEL file
worldseries

26. **World Series in Major League Baseball.** Major League Baseball's World Series is a maximum of seven games, with the winner being the first team to win four games. Assume that the Atlanta Braves are playing the Minnesota Twins in the World Series and that the first two games are to be played in Atlanta, the next three games at the Twins' ballpark, and the last two games, if necessary, back in Atlanta. Taking into account the projected starting pitchers for each game and the home field advantage, the probabilities of Atlanta winning each game are listed in the following table. LO 3

Game	1	2	3	4	5	6	7
Probability of Atlanta Win	0.60	0.55	0.48	0.45	0.48	0.55	0.50

a. Set up a spreadsheet simulation model in which the outcome of each game (whether Atlanta or Minnesota wins) is a random variable.
b. What is the average number of games played regardless of the winner?
c. What is the probability that the Atlanta Braves win the World Series?

MODEL file
anime

27. **Newsvendor Problem for Anime Magazine.** Young entrepreneur Susie Perez has launched a business venture in which Susie uses stories submitted by university students as the basis for comics in a monthly anime-style magazine. Based on market research, Susie estimates that average monthly demand will be 500 copies. Susie has decided to model monthly demand as normal random variable with a mean of 500 and a standard deviation of 300.

Susie must pay a publishing company $3.75 for each copy of the comic printed. Susie then sells the magazine for $5 each. Rather than having a store-front, Susie sells the magazines through a group of student vendors who sell the comics out of their backpacks while on campus. Susie pays a student vendor $0.35 for each magazine sold. Susie only sells each issue for one month, and then a new issue is sold. However, the publishing company has agreed to buy back any unsold copies at the end of each month for $2.25. Construct a simulation model that computes Susie's profit. LO 2

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. As Susie validates the simulation model you have constructed, Susie observes that when using a normal distribution with a mean of 500 and a standard deviation of 300 to model demand, negative values of demand are often generated. How can you modify the simulation model to address this issue?
b. Based on the simulation model that incorporates a remedy for the validation issue observed in part (a), what is the estimate of the average profit if Susie sets the order quantity to 1,200? What is the 95% confidence interval on this estimate of average profit?
c. For an order quantity of 1,200 copies, what is the profit value such that 2.5% of the profit outcomes are smaller than this value? What is the profit value such that 2.5% of the profit outcomes are larger than this value? Why doesn't this range correspond to the 95% confidence interval in part (b)?

MODEL file
salestrips

28. **Product Adoption and Sales Bonus.** Bianca Peterson is a marketing engineer for Hexagon Composites, a company which sells carbon composite storage tanks. In an effort to gain product adoptions from customers, Bianca goes on sales trips, often to foreign countries. For each of 120 previous sales trips, the file *salestrips* lists (1) whether the trip resulted in the visited customer adopting the product, and (2) the revenue generated by the adoption. **LO 2, 3**
a. Bianca has six sales trips planned over the next couple of months. Construct a simulation model to compute the total revenue generated from these six trips.
b. What is the average revenue that Bianca expects to generate from the six trips?
c. Bianca earns a bonus if more than $500,000 in revenue is generated on the six trips. What is the probability that Bianca earns the bonus?

MODEL file
pancheros

29. **Financial Analysis of Restaurant.** Pancheros sells a variety of Mexican-inspired cuisine for which tortillas are often a main ingredient. Assume that each customer places an order requiring one tortilla with a 75% probability independent of other customers' orders. The other 25% of customers place orders that do not require a tortilla. Assume that the number of customers who arrive per hour has a Poisson distribution with the average number of customers in an hour time slot given in the following table:

Time of Day	Average Number of Customers
11am–12pm	200
12pm–1pm	200
1pm–2pm	200
2pm–3pm	50
3pm–4pm	50
4pm–5pm	50
5pm–6pm	150
6pm–7pm	150
7pm–8pm	150
8pm–9pm	50
9pm–10pm	50

Pancheros currently prepares dough for 750 tortillas at the beginning of each day. Due to uncertain customer demand, Pancheros may run out of tortillas, which affects profit as well as customer relations. Every tortilla-based customer order generates $2.35 in profit. Every customer who places an order requiring a tortilla but is denied (due to a tortilla stock-out) leaves Pancheros without buying anything with probability 0.13, and purchases a non-tortilla menu item (generating profit of $1.50) with probability 0.87. **LO 3**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Create a simulation model to generate the distribution of daily lost profit due to tortilla stock-outs.
b. What is the average daily lost profit? What is the 95% confidence interval on this mean?
c. On average, which hour of the work day does Pancheros run out of tortillas? What is the 95% confidence interval on the mean?

MODEL file
admissions

30. **Matriculation Yield at a University.** As admissions director for an exclusive executive MBA program which takes place on Necker Island in the Caribbean, Richard Branson must decide which applicants should receive admission offers. This is a difficult decision-making problem, as an applicant may or may not accept an admission offer. Currently, Richard is considering 30 applicants, each of which has a different probability of accepting an admission offer. Based on their academic qualifications and experience, Richard has rated each of these applicants using a value score from 1 to 10 (higher value scores represent better applicants). The file *admissions* contains data on the 30 applicants.

Based on the capacity of their facilities, Richard would like a class of 12 students. Fewer students than 12 results in under-utilized resources (empty classroom seats), but more than 12 students results in increased marginal costs. Specifically, each attending student beyond 12 incurs a cost of 20 value points. Note that an applicant will be an attending student only if they are admitted and they accept the admission offer. **LO 3**

a. Construct a spreadsheet model that computes the net value of offering admission to the top 20 students as ranked by value score.
b. Compute net value as the sum of the value of attending students minus the costs of students beyond the capacity of 12 seats. What is the average net value obtained when offering admission to the top 20 students?

MODEL file
wedding

31. **Wedding Attendance.** The wedding date for a couple is quickly approaching, and the wedding planner must provide the caterer an estimate of how many people will attend the reception so that the appropriate quantity of food is prepared for the buffet. The following table contains information on the number of RSVPs for the 145 invitations. Unfortunately, the number of guests who actually attend does not always correspond to the number of RSVPs.

Based on the planner's experience, the wedding planner knows that it is extremely rare for guests to attend a wedding if they affirmed that they will not be attending. Therefore, the wedding planner will assume that no one from these 50 invitations will attend. The wedding planner estimates that each of the 25 guests planning to come alone has a 75% chance of attending alone, a 20% chance of not attending, and a 5% chance of bringing a companion. For each of the 60 RSVPs who plan to bring a companion, there is a 90% chance that they will attend with a companion, a 5% chance of attending alone, and a 5% chance of not attending at all. For the 10 people who have not responded, the wedding planner assumes that there is an 80% chance that each will not attend, a 15% chance that they will attend alone, and a 5% chance that they will attend with a companion. **LO 3**

RSVPs	No. of Invitations
0	50
1	25
2	60
No response	10

a. Assist the wedding planner by constructing a spreadsheet simulation model to estimate the average number of guests who will attend the reception.
b. To be accommodating hosts, the couple has instructed the wedding planner to use the simulation model to determine X, the minimum number of guests for which the caterer should prepare the meal, so that there is at least a 90% chance that the actual attendance is less than or equal to X. What is the best estimate for the value of X?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

MODEL file
quarterlyexchange

32. **Hedging Currency Risk.** A European put option on a currency allows you to sell a unit of that currency at the specified strike price (exchange rate) at a particular point in time after the purchase of the option. For example, suppose Press Teag Worldwide (from Section 13.4) purchases a three-month European put option for a British pound with a strike price of £0.630 per U.S. dollar. Then, if the exchange rate in three months is such that it takes more than £0.630 to buy a U.S. dollar, for example, £0.650 per U.S. dollar, Press Teag will exercise the put option and sell its pound sterling at the strike price of £0.630 per U.S. dollar. However, if exchange rate in three months is such that it take less than £0.630 to buy a U.S. dollar, for example, £0.620 per U.S. dollar, Press Teag will not exercise its put option and sell its pound sterling at the market rate of £0.620 per U.S. dollar.

The following table lists information on the three-month European put options on pound sterling, New Zealand dollars, and Japanese yen.

Currency	Purchase Price per Option	Strike Price
Pound Sterling	$0.01 per £	£0.630 per $
New Zealand Dollar	$0.01 per NZD	NZD 1.230 per $
Japanese Yen	$0.00005 per ¥	¥90.00 per $

a. Modify the simulation model in the file *quarterlyexchange* to compare the strategy of hedging half of the revenue in each of three foreign currencies using European put options versus the strategy of not using put options to hedge at all. **LO 4, 5**
b. What is the average difference in revenue (hedged revenue – unhedged revenue)?
c. What is the probability that the hedged revenue is less than the unhedged revenue?

MODEL file
dailystock

33. **Modeling Stock Prices.** Over the past year, a financial analyst has tracked the daily change in the price per share of common stock for a major oil company. The financial analyst wants to develop a simulation model to analyze the stock price at the end of the next quarter. Assume 63 trading days and a current price per share of $51.60. **LO 2, 4, 5**
a. Based on the data in the *DataToFit* worksheet of the file *dailystock*, use the Excel formula =CORREL(B3:B313, B4:B314) to compute the correlation between the percent change in stock price on consecutive days. What do you conclude about the dependency of the percent change in stock price from day to day?
b. Based on the data in the *DataToFit* worksheet of the file *dailystock*, compute sample statistics and construct a histogram to visualize the distribution of the data. Select a distribution that appears to fit this data. *Hint:* Review Appendix 13.1 for descriptions of various types of probability distributions.
c. Using the distribution that you selected in part (b) to represent the daily percent change in stock price, construct a simulation model to estimate the price per share at the end of the quarter. What is the probability that the stock price will be below $26.55?
d. The *WhatReallyHappened* worksheet of the file *dailystock* contains the 63 values of the daily percent change in stock price that actually occurred during the quarter. What does this reveal about the limitations of simulation modeling? What could the financial analyst do to address this limitation?

MODEL file
burgerdome

34. **Restaurant Wait Time.** Burger Dome is a fast-food restaurant currently evaluating its customer service. In its current operation, an employee takes a customer's order, tabulates the cost, receives payment from the customer, and then fills the order. Once the customer's order is filled, the employee takes the order of the next customer waiting for service. Assume that time between each customer's arrival is an exponential random variable with a mean of 1.35 minutes. Assume that the time for the employee to complete the customer's service is an exponential random variable with a mean of 1 minute. Use the file *burgerdome* to complete a simulation model for the waiting line at Burger Dome for a 14-hour workday. Using the summary statistics gathered at the bottom of the spreadsheet model, answer the following questions. **LO 2**
a. What is the average wait time experienced by a customer?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. What is the longest wait time experienced by a customer?
c. What is the probability that a customer waits more than 2 minutes?
d. Create a histogram depicting the wait time distribution with bins of minutes corresponding to [0,1], (1,2], (2,3], and so on. What is the most common wait time bin?
e. By pressing the F9 key to generate a new set of simulation trials, you can observe the variability in the summary statistics from simulation to simulation. Typically, this variability can be reduced by increasing the number of trials. Why is this approach not appropriate for this problem?

35. **Effect of Service Time Distribution on Restaurant Wait Time.** One advantage of simulation is that a simulation model can be altered easily to reflect a change in the assumptions. Refer to the Burger Dome analysis in Problem 34. Assume that the service time is more accurately described by a normal distribution with a mean of 1 minute and a standard deviation of 0.2 minute. This distribution has less variability than the exponential distribution originally used. What is the impact of this change on the output measures? **LO 2**

MODEL file
burgerdome2

36. **Effect of Additional Server on Restaurant Wait Time.** Refer to the Burger Dome analysis in Problem 34. Burger Dome wants to consider the effect of hiring a second employee to serve customers (in parallel with the first employee). Use the file *burgerdome2* to complete a simulation model that accounts for the second employee. What is the impact of this change on the output measures? **LO 2**

Case Problem 1: Four Corners

What will your investment portfolio be worth in 10 years? In 20 years? When you stop working? The Human Resources Department at Four Corners Corporation was asked to develop a financial planning model that would help employees address these questions. Tom Gifford was asked to lead this effort and decided to begin by developing a personal financial plan. Tom has a degree in business and, at the age of 40, is making $85,000 per year. Through contributions to the company's retirement program and the receipt of a small inheritance, Tom has accumulated a portfolio valued at $50,000. Tom plans to work 20 more years and hopes to accumulate a portfolio valued at $1,000,000. Can Tom do it?

Tom began with a few assumptions about future salary, new investment contributions, and the portfolio growth rate. Tom assumed a 5% annual salary growth rate and plans to make new investment contributions at 6% of the annual salary. After some research on historical stock market performance, Tom decided that a 10% annual portfolio growth rate was reasonable. Using these assumptions, Tom developed the following Excel worksheet:

The worksheet provides a financial projection for the next five years. In computing the portfolio earnings for a given year, Tom assumed that their new investment contribution

MODEL file
fourcorners

	A	B	C	D	E	F	G
1	**Four Corners**						
2							
3	Age	40					
4	Current Salary	$85,000					
5	Current Portfolio	$50,000					
6	Annual Investment Rate	6%					
7	Salary Growth Rate	5%					
8	Portfolio Growth Rate	10%					
9							
10	Year	Beginning Balance	Salary	New Investment	Earnings	Ending Balance	Age
11	1	$50,000	$85,000	$5,100	$5,255	$60,355	41
12	2	$60,355	$89,250	$5,355	$6,303	$72,013	42
13	3	$72,013	$93,713	$5,623	$7,482	$85,118	43
14	4	$85,118	$98,398	$5,904	$8,807	$99,829	44
15	5	$99,829	$103,318	$6,199	$10,293	$116,321	45
16							

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

would occur evenly throughout the year, and thus half of the new investment could be included in the computation of the portfolio earnings for the year. From the worksheet, we see that, at age 45, Tom is projected to have a portfolio valued at $116,321.

Tom's plan was to use this worksheet as a template to develop financial plans for the company's employees. The data in the spreadsheet would be tailored for each employee, and rows would be added to it to reflect the employee's planning horizon. After adding another 15 rows to the worksheet, Tom found that one could expect to have a portfolio of $772,722 after 20 years. Tom then took the results to show senior human resources manager, Kate Krystkowiak.

Although Kate was pleased with Tom's progress, there was still room for improvement. One of the criticisms was the assumption of a constant annual salary growth rate. Most employees experience some variation in the annual salary growth rate from year to year. In addition, the constant annual portfolio growth rate was unrealistic and that the actual growth rate would vary considerably from year to year. Kate and Tom both agreed that a simulation model for the portfolio projection would allow them to account for the random variability in the salary growth rate and the portfolio growth rate.

After some research, Tom and Kate decided to assume that the annual salary growth rate would vary from 0% to 5% and that a uniform probability distribution would provide a realistic approximation. Four Corners' accountants suggested that the annual portfolio growth rate could be approximated by a normal probability distribution with a mean of 10% and a standard deviation of 5%. With this information, Tom set off to redesign the spreadsheet so that it could be used by the company's employees for financial planning. **LO 2**

Managerial Report

Play the role of Tom Gifford, and develop a simulation model for financial planning. Write a report for Tom's boss and, at a minimum, include the following:

In Chapter 12, we discuss the details of Excel's Goal Seek function.

1. Without considering the random variability, extend the current worksheet to 20 years. Confirm that by using the constant annual salary growth rate and the constant annual portfolio growth rate, Tom can expect to have a 20-year portfolio of $772,722. What would Tom's annual investment rate have to increase to in order for the portfolio to reach a 20-year, $1,000,000 goal? *Hint:* Use Goal Seek in Excel.
2. Redesign the spreadsheet model to incorporate the random variability of the annual salary growth rate and the annual portfolio growth rate into a simulation model. Assume that Tom is willing to use the annual investment rate that predicted a 20-year, $1,000,000 portfolio in part (1). Show how to simulate Tom's 20-year financial plan. Use results from the simulation model to comment on the uncertainty associated with Tom reaching the 20-year, $1,000,000 goal.
3. What recommendations do you have for employees with a current profile similar to Tom's after seeing the impact of the uncertainty in the annual salary growth rate and the annual portfolio growth rate?
4. Assume that Tom is willing to consider working 25 more years instead of 20 years. What is your assessment of this strategy if Tom's goal is to have a portfolio worth $1,000,000?
5. Discuss how the financial planning model developed for Tom Gifford can be used as a template to develop a financial plan for any of the company's employees.

Case Problem 2: Ginsberg's Jewelry Snowfall Promotion

To boost holiday sales, Ginsberg's Jewelry is contemplating a sales promotion for the store. Under the promotion, any jewelry purchase made between Thanksgiving and Christmas is eligible for a discount that will be determined by the snowfall in the first seven days of the new year (January 1–7). Specifically, the purchase discount percentage will be computed using the formula

$$\max\left[(\textit{cumulative snowfall over first seven days of January} \div 20) \times 100\%, 100\%\right]$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The max function in the formula is because Ginsberg states in the fine print of the promotion that the maximum rebate is 100% (in case there is more than 20 inches of snowfall in the first seven days of January). For example, if there is a total of 5 inches of snow from January 1 to January 7, any customer who made a purchase between Thanksgiving and Christmas of the previous year can return to the store and claim a rebate of max [(5 ÷ 20) × 100%,100%] = 25% of their original purchase price.

DATA *file* snowfall

Under ordinary circumstances, Ginsberg's total sales between Thanksgiving and Christmas are normally distributed with a mean of $300,000 and a standard deviation of $35,000. Ginsberg believes that the promotion has the potential for boosting sales anywhere from 0% to 20%, but does not have any further information on this uncertain impact of the promotion. Ginsberg's gross margin (without any sales promotions) is 42%.

Ashley Ginsberg, the niece of the proprietor of Ginsberg's Jewelry, is planning to construct a simulation model that compares the gross profit with the proposed promotion to the gross profit without the promotion. As a first step, Ashley has collected historical daily snowfall amounts for the area. After examining the data, Ashley has determined that daily historical snowfall in the months of December, January, and February over an 81-year period is representative of the daily snowfall from January 1 to January 7 in any particular year. Wrangling the data results in a data set of 7,212 daily snowfall amounts listed in the *Data* worksheet of the file *snowfall*. **LO 2, 4, 5**

Data wrangling is discussed in Chapter 4.

Managerial Report

1. Create a what-if model that computes the difference between gross profit with the promotion and gross profit without the promotion. In the construction of this what-if model, assume that sales without the promotion are $300,000 and that the promotion boosts sales by 10%, and the snowfall on the seven days is given by the sequence {0, 0, 1.75, 0.25, 0, 0, 0}. For these values, is the promotion a financial success?
2. As a next step in explicitly accounting for the uncertainty in daily snowfall, examine the snowfall data to determine if there is any dependency between the snowfall on consecutive days.
   a. First, compute the correlation between snowfall on consecutive days to determine if a linear dependency exists.[1] What does this suggest?
   b. Now examine a different type of dependency by using a PivotTable to estimate the following:
      i. $P(\text{Snowfall} > 0 \text{ tomorrow} \mid \text{Snowfall} = 0 \text{ today})$ and $P(\text{Snowfall} = 0 \text{ tomorrow} \mid \text{Snowfall} = 0 \text{ today})$
      ii. $P(\text{Snowfall} > 0 \text{ tomorrow} \mid \text{Snowfall} > 0 \text{ today})$ and $P(\text{Snowfall} = 0 \text{ tomorrow} \mid \text{Snowfall} > 0 \text{ today})$
      iii. What do the calculations in parts (b-i) and (b-ii) suggest?
3. Modify the what if-model from part 1 by explicitly modeling the following quantities as random variables:
   a. Sales amount without promotion
   b. Percentage that promotion boosts sales
   c. Daily snowfall amount on each of seven days from January 1 to January 7

   In particular, be sure to model daily snowfall amount appropriately in light of your observations in part (2).
4. What is a 95% confidence interval on the difference in gross profit with and without the snowfall promotion?
5. What is a 95% confidence interval on the probability the snowfall promotion is a financial success?

[1]Careful inspection of the snowfall data will reveal that not all consecutive observations in the data correspond to consecutive days due to the data wrangling (and using only data from December, January, and February of each year). However, this only affects a small portion of the data. Proceed with analysis of the data assuming all data is temporally consecutive as this only introduces slight error in the analysis.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 13 Appendix

Appendix 13.1 Common Probability Distributions for Simulation

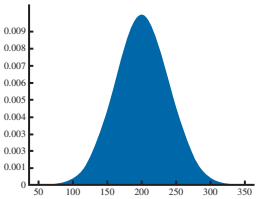
Simulation software such as Analytic Solver, Crystal Ball, or @RISK automates the generation of random values from an even wider selection of probability distributions than is available in native Excel.

Selecting the appropriate probability distribution to characterize a random variable in a simulation model can be a critical modeling decision. In this appendix, we review several probability distributions commonly used in simulation models. We describe the native Excel functionality used to generate random values from the corresponding probability distribution.

Continuous Probability Distributions

Random variables that can be many possible values (even if the values are discrete) are often modeled with a continuous probability distribution. For common continuous random variables, we provide several pieces of information. First, we list the parameters which specify the probability distribution. We then delineate the minimum and maximum values defining the range that can be realized by a random variable that follows the given distribution. We also provide a short description of the overall shape of the distribution paired with an illustration. Then, we supply an example of the application of the random variable. We conclude with the native Excel functionality for generating random values from the probability distribution.

Normal Distribution



Parameters: mean (m), standard deviation (s)

Range: $-\infty$ to $+\infty$

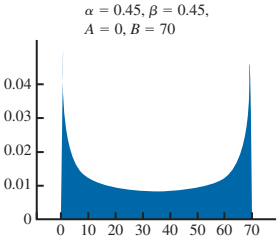
Description: The normal distribution is a bell-shaped, symmetric distribution centered at its mean m. The normal distribution is often a good way to characterize a quantity that is the sum of many independent random variables.

Example: In human resource management, employee performance is often well represented by a normal distribution. Typically, the performance of 68% of employees is within one standard deviation of the average performance, and the performance of 95% of the employees is within two standard deviations. Employees with exceptionally low or high performance are rare. For example, the performance of a pharmaceutical company's sales force may be well described by a normal distribution with a mean of 200 customer adoptions and a standard deviation of 40 customer adoptions.

Native Excel: NORM.INV(RAND(), m, s)

Beta Distribution

MODEL *file*
beta



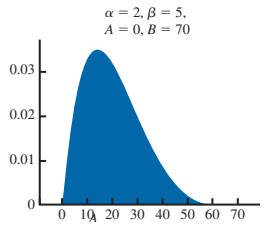
Parameters: alpha (α), beta (β), minimum (A), maximum (B)

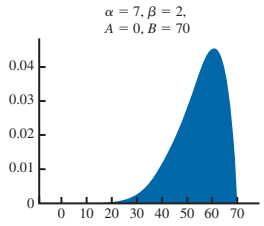
Range: A to B

Description: Over the range specified by values A and B, the beta distribution has a very flexible shape that can be manipulated by adjusting α and β. The beta distribution is useful in modeling an uncertain quantity that has a known minimum and maximum value. To estimate the values of the alpha and beta parameters from sample data, we use the following equations (see the file *beta* for an Excel implementation):

$$\alpha = \left(\frac{\bar{x} - A}{B - A}\right)\left(\frac{\left(\frac{\bar{x} - A}{B - A}\right)\left(1 - \left(\frac{\bar{x} - A}{B - A}\right)\right)}{\frac{s^2}{(B - A)^2}} - 1\right)$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.





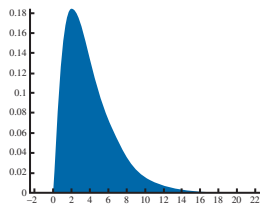
$$\beta = \alpha \times \left(\frac{\left(1 - \left(\frac{\overline{x} - A}{B - A}\right)\right)}{\left(\frac{\overline{x} - A}{B - A}\right)} \right)$$

Example: The boom-or-bust nature of the revenue generated by a movie from a polarizing director may be described by a beta distribution. The relevant values (in millions of dollars) are $A = 0$, $B = 70$, $\alpha = 0.45$, and $\beta = 0.45$. This particular distribution is U-shaped and extreme values are more likely than moderate values. The figures in the left margin illustrate beta distributions with different values of α and β, demonstrating its flexibility. The first figure depicts a U-shaped beta distribution. The second figure depicts a unimodal beta distribution with a positive skew. The third figure depicts a unimodal beta distribution with a negative skew.

Native Excel: BETA.INV(RAND(), α, β, A, B)

Gamma Distribution

MODEL file
gamma



Parameters: alpha (α), beta (β)

Range: 0 to $+\infty$

Description: The gamma distribution has a very flexible shape controlled by the values of α and β. The gamma distribution is useful in modeling an uncertain quantity that can be as small as zero but can also realize large values. To estimate the values of the alpha and beta parameters given sample data, we use the following equations (see the file *gamma* for an Excel implementation):

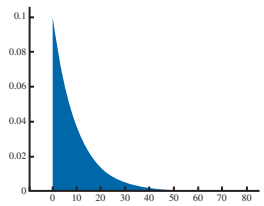
$$\alpha = \left(\frac{\overline{x}}{s}\right)^2$$

$$\beta = \frac{s^2}{\overline{x}}$$

Example: The aggregate amount (in $100,000s) of insurance claims in a region may be described by a gamma distribution with $\alpha = 2$ and $\beta = 0.5$.

Native Excel: GAMMA.INV(RAND(), α, β)

Exponential Distribution



Parameters: mean (m)

Range: 0 to $+\infty$

Description: The exponential distribution is characterized by a mean value equal to its standard deviation and a long right tail stretching from a mode value of 0.

Example: The time between events, such as customer arrivals or customer defaults on bill payment, are commonly modeled with an exponential distribution. An exponential random variable possesses the "memoryless" property: the probability of a customer arrival occurring in the next x minutes does not depend on how long it's been since the last arrival. For example, suppose the average time between customer arrivals is 10 minutes. Then, the probability that there will be 25 or more minutes between customer arrivals if 10 minutes have passed since the last customer arrival is the same as the probability that there will be more than 15 minutes until the next arrival if a customer just arrived.

Native Excel: LN(1-RAND())*(−m), or equivalently GAMMA.INV(RAND(), 1, m) because the exponential distribution is a special case of the gamma distribution with $\alpha = 1$ and $\beta = m$

Triangular Distribution

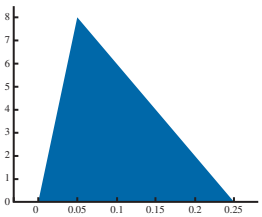
MODEL file
triangular

Parameters: minimum (a), most likely (m), maximum (b)

Range: a to b

Description: The triangular distribution is often used to subjectively assess uncertainty when little is known about a random variable besides its range, but it is thought to have a single mode. The distribution is shaped like a triangle with vertices at a, m, and b.

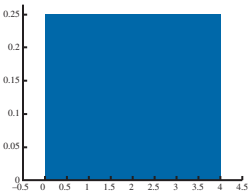
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Example: In corporate finance, a triangular distribution may be used to model a project's annual revenue growth in a net present value analysis if the analyst can reliably provide minimum, most likely, and maximum estimates of growth. For example, a project may have worst-case annual revenue growth of 0%, a most-likely annual revenue growth of 5%, and best-case annual revenue growth of 25%. These values would then serve as the parameters for a triangular distribution.

Native Excel: IF(*random* $<$ $(m-a)/(b-a)$, a + SQRT($(b-a)$*$(m-a)$**random*), b − SQRT($(b-a)$*$(b-m)$*$(1-$ *random*))), where *random* refers to a single, separate cell containing =RAND() (see the file *triangular* for an Excel implementation)

Uniform Distribution



Parameters: minimum (a), maximum (b)

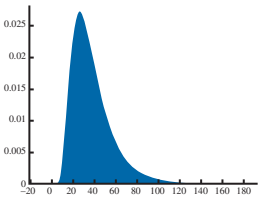
Range: a to b

Description: The uniform distribution is appropriate when a random variable is equally likely to be any value between a and b. When little is known about a phenomenon other than its minimum and maximum possible values, the uniform distribution may be a conservative choice to model an uncertain quantity.

Example: A service technician making a house call may quote a 4-hour time window in which the technician will arrive. If the technician is equally likely to arrive any time during this time window, then the arrival time of the technician in this time window may be described with a uniform distribution.

Native Excel: $a + (b-a)*$RAND()

Log-Normal Distribution



Parameters: *log_mean*, *log_stdev*

Range: 0 to $+\infty$

Description: The log-normal distribution is a unimodal distribution (like the normal distribution) that has a minimum value of 0 and a long right tail (unlike the normal distribution). The log-normal distribution is often a good way to characterize a quantity that is the product of many independent, positive random variables. The natural logarithm of a log-normally distributed random variable is normally distributed.

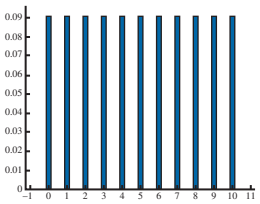
Example: The income distribution of the lower 99% of a population is often well described using a log-normal distribution. For example, for a population in which the natural logarithm of the income observations is normally distributed with a mean of 3.5 and a standard deviation of 0.5, the income observations are distributed log-normally.

Native Excel: LOGNORM.INV(RAND(), *log_mean*, *log_stdev*), where *log_mean* and *log_stdev* are the mean and standard deviation of the normally distributed random variable obtained when taking the logarithm of the log-normally distributed random variable.

Discrete Probability Distributions

Random variables that can be only a relatively small number of discrete values are often best modeled with a discrete distribution. The appropriate choice of discrete distribution relies on the specific situation. For common discrete random variables, we provide several pieces of information. First, we list the parameters required to specify the distribution. Then, we outline possible values that can be realized by a random variable that follows the given distribution. We also provide a short description of the distribution paired with an illustration. Then, we supply an example of the application of the random variable. We conclude with the native Excel functionality for generating random values from the probability distribution.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



Integer Uniform Distribution

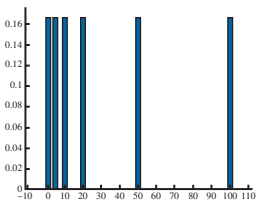
Parameters: lower (l), upper (u)

Possible values: $l, l + 1, l + 2, \ldots, u - 2, u - 1, u$

Description: An integer uniform random variable assumes that the integer values between l and u are equally likely.

Example: The number of philanthropy volunteers from a class of 10 students may be an integer uniform variable with values 0, 1, 2, . . . , 10.

Native Excel: RANDBETWEEN(l, u)



Discrete Uniform Distribution

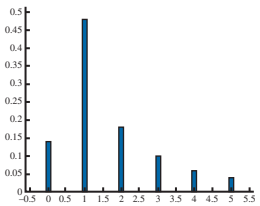
Parameters: set of values $\{v_1, v_2, v_3, \ldots, v_k\}$

Possible values: $v_1, v_2, v_3, \ldots, v_k$

Description: A discrete uniform random variable is equally likely to be any of the specified set of values $\{v_1, v_2, v_3, \ldots, v_k\}$.

Example: Consider a game show that awards a contestant a cash prize from an envelope randomly selected from six possible envelopes. If the envelopes contain \$1, \$5, \$10, \$20, \$50, and \$100, respectively, then the prize is a discrete uniform random variable with values {1, 5, 10, 20, 50, 100}.

Native Excel: CHOOSE(RANDBETWEEN(*1, k*), $v_1, v_2, \ldots, v_k$)



Custom Discrete Distribution

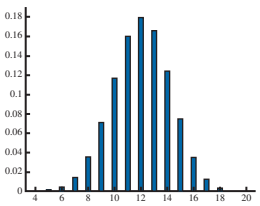
Parameters: set of values $\{v_1, v_2, v_3, \ldots, v_k\}$ and corresponding weights $\{w_1, w_2, w_3, \ldots, w_k\}$ such that $w_1 + w_2 + \cdots + w_k = 1$

Possible values: $v_1, v_2, v_3, \ldots, v_k$

Description: A custom discrete distribution can be used to create a tailored distribution to model a discrete, uncertain quantity. The value of a custom discrete random variable is equal to the value v_i with probability w_i.

Example: Analysis of daily sales for the past 50 days at a car dealership shows that on 7 days no cars were sold, on 24 days one car was sold, on 9 days two cars were sold, on 5 days three cars were sold, on 3 days four cars were sold, and on 2 days five cars were sold. We can estimate the probability distribution of daily sales using the relative frequencies. An estimate of the probability that no cars are sold on a given day is $7/50 = 0.14$, an estimate of the probability that one car is sold is $24/50 = 0.48$, and so on. Daily sales may then be described by a custom discrete distribution with values of {0, 1, 2, 3, 4, 5} with respective weights of {0.14, 0.48, 0.18, 0.10, 0.06, 0.04}.

Native Excel: First, create a table with three columns so that the first column lists each possible value of the uncertain quantity, the second column lists the probability of the corresponding possible value, and the third column computes the cumulative probability corresponding to the sum of the current row's probability and all previous rows' probabilities. Second, use the RAND function and a XLOOKUP function to return the value corresponding to a probability generated by the RAND function. Figure 13.33 illustrates the implementation in *customdiscrete* for the car sales example.



Binomial Distribution

Parameters: trials (n), probability of a success (p)

Possible values: 0, 1, 2, . . . , n

Description: A binomial random variable corresponds to the number of times an event successfully occurs in n trials, and the probability of a success at each trial is p and independent of whether a success occurs on other trials. When $n = 1$, the binomial is also known as the Bernoulli distribution.

Example: In a portfolio of 20 similar stocks, each of which has the same probability of increasing in value of $p = 0.6$, the total number of stocks that increase in value can be described by a binomial distribution with parameters $n = 20$ and $p = 0.6$.

Native Excel: BINOM.INV(n, p, RAND())

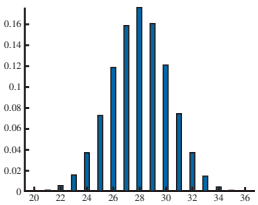
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.33 Native Excel Implementation of Custom Discrete Distribution

	A	B	C
1			Random Number
2	Cars Sold	=XLOOKUP(C2,C5:C10,A5:A10,"error",1,2)	=RAND()
3			
4	Possible Values	Probability	Cumulative Probability
5	0	0.14	=SUM($B$5:B5)
6	1	0.48	=SUM($B$5:B6)
7	2	0.18	=SUM($B$5:B7)
8	3	0.1	=SUM($B$5:B8)
9	4	0.06	=SUM($B$5:B9)
10	5	0.04	=SUM($B$5:B10)

MODEL *file*
customdiscrete

	A	B	C
1			Random Number
2	Cars Sold	1	0.5226
3			
4	Possible Values	Probability	Cumulative Probability
5	0	0.14	0.14
6	1	0.48	0.62
7	2	0.18	0.80
8	3	0.10	0.90
9	4	0.06	0.96
10	5	0.04	1.00



Hypergeometric Distribution

Parameters: trials (n), population size (N), successful elements in population (s)

Possible values: $\max\{0, n + s - N\}, \ldots, \min\{n, s\}$

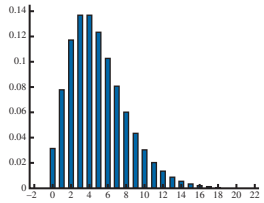
Description: A hypergeometric random variable corresponds to the number of times an element labeled a success is selected out of n trials in the situation where there are N total elements, s of which are labeled a success and, once selected, cannot be selected again. Note that this is similar to the binomial distribution except that now the trials are dependent because removing the selected element changes the probabilities of selecting an element labeled a success on subsequent trials.

Example: A certain company produces circuit boards to sell to computer manufacturers. Because of a quality defect in the manufacturing process, it is known that only 70 circuit boards out of a lot of 100 have been produced correctly and the other 30 are faulty. If a company orders 40 circuit boards from this lot of 100, the number of functioning circuit boards that the company will receive in their order is a hypergeometric random variable with $n = 40$, $s = 70$, and $N = 100$. Note that, in this case, between 10 ($= 40 + 70 - 100$) and 40 ($=\min\{40, 70\}$) of the 40 ordered circuit boards will be functioning. At least 10 of the 40 circuit boards will be functioning because at most 30 ($= 100 - 70$) are faulty.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Native Excel: Insert the worksheet from the file *hypergeometric* into your Excel workbook, modify the parameters in the cell range B2:B4, and then reference cell B6 in your simulation model to obtain a value from a hypergeometric distribution. This file uses the RAND function in conjunction with the XLOOKUP function referencing a table in which each row lists a possible value and the cumulative probability associated with the corresponding value; the cumulative probability of each value is computed with the HYPGEOM.DIST function. Figure 13.34 illustrates *hypergeometric* with the parameter values should only be a single space for the circuit board example.

Negative Binomial Distribution



Parameters: required number of successes (s), probability of success (p)

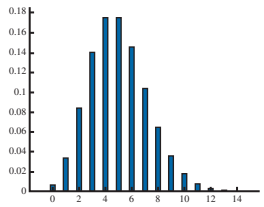
Possible values: 0, 1, 2, . . . , ∞

Description: A negative binomial random variable corresponds to the number of times that an event fails to occur until an event successfully occurs s times, given that the probability of an event successfully occurring at each trial is p. When $s = 1$, the negative binomial is also known as the geometric distribution.

Example: Consider the research and development (R&D) division of a large company. An R&D division may invest in several projects that fail before investing in 5 projects that succeed. If each project has a probability of success of *0.50*, the number of projects that fail before 5 successful projects occur is a negative binomial random variable with parameters $s = 5$ and $p = 0.50$.

Native Excel: Insert the worksheet from the file *negativebinomial* into your Excel workbook, modify the parameters in the cell range B2:B3, and then reference cell B5 in your simulation model to obtain a value from a negative binomial distribution. This file uses the RAND function in conjunction with the XLOOKUP function referencing a table in which each row lists a possible value and the cumulative probability associated with the corresponding value; the cumulative probability of each value is computed with the NEGBINOM.DIST function. Figure 13.35 illustrates the *negativebinomial* with the parameter values for the R&D project example.

Poisson Distribution



Parameters: mean (m)

Possible values: 0, 1, 2, . . .

Description: A Poisson random variable corresponds to the number of times that an event occurs within a specified period of time given that m is the average number of events within the specified period of time.

Example: The number of patients arriving at a health care clinic in an hour can be modeled with a Poisson random variable with $m = 5$, if on average 5 customers arrive to the store in an hour.

Native Excel: Insert the worksheet from the file *poisson* into your Excel workbook, modify the parameter in cell B2, and then reference cell B4 in your simulation model to obtain a value from a Poisson distribution. This file uses the RAND function in conjunction with the XLOOKUP function referencing a table in which each row lists a possible value and the cumulative probability associated with the corresponding value; the cumulative probability of each value is computed with the POISSON.DIST function. Figure 13.36 illustrates the *poisson* with the parameter value for the health care clinic example.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.34 Excel Template to Generate Values from a Hypergeometric Distribution

	A	B	C
1	**Hypergeometric Distribution Parameters**		
2	Trials (n)	40	
3	Population (N)	100	
4	Successful Elements in Population (s)	70	
5			Random Number
6	Randomly Generated Hypergeometric Value	=XLOOKUP(C6,B9:B109,A9:A109,"error",1,2)	=RAND()
7			
8	Number of Successes in n Trials	Cumulative Probability	
9	0	=HYPGEOM.DIST(A9,$B$2,$B$4,$B$3,1)	
10	1	=HYPGEOM.DIST(A10,$B$2,$B$4,$B$3,1)	
11	2	=HYPGEOM.DIST(A11,$B$2,$B$4,$B$3,1)	
12	3	=HYPGEOM.DIST(A12,$B$2,$B$4,$B$3,1)	
13	4	=HYPGEOM.DIST(A13,$B$2,$B$4,$B$3,1)	
14	5	=HYPGEOM.DIST(A14,$B$2,$B$4,$B$3,1)	
15	6	=HYPGEOM.DIST(A15,$B$2,$B$4,$B$3,1)	
16	7	=HYPGEOM.DIST(A16,$B$2,$B$4,$B$3,1)	
17	8	=HYPGEOM.DIST(A17,$B$2,$B$4,$B$3,1)	
18	9	=HYPGEOM.DIST(A18,$B$2,$B$4,$B$3,1)	
19	10	=HYPGEOM.DIST(A19,$B$2,$B$4,$B$3,1)	

	A	B	C
1	**Hypergeometric Distribution Parameters**		
2	Trials (n)	40	
3	Population (N)	100	
4	Successful Elements in Population (s)	70	
5			Random Number
6	Randomly Generated Hypergeometric Value	24	0.0476
7			
8	Number of Successes in n Trials	Cumulative Probability	
9	0	0.000	
10	1	0.000	
11	2	0.000	
12	3	0.000	
13	4	0.000	
14	5	0.000	
15	6	0.000	
16	7	0.000	
17	8	0.000	
18	9	0.000	
19	10	0.000	
20	11	0.000	
21	12	0.000	
22	13	0.000	
23	14	0.000	
24	15	0.000	
25	16	0.000	
26	17	0.000	
27	18	0.000	
28	19	0.000	
29	20	0.000	
30	21	0.002	
31	22	0.007	
32	23	0.023	
33	24	0.060	
34	25	0.133	

MODEL *file*

hypergeometric

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.35 Excel Template to Generate Values from a Negative Binomial Distribution

	A	B	C
1	**Negative Binomial Distribution Parameters**		
2	Required Number of Successes (s)	5	
3	Probability of Success (p)	0.5	
4			Random Number
5	Randomly Generated Negative Binomial Value	=XLOOKUP(C5,B8:B508,A8:A508,"error",1,2)	=RAND()
6			
7	Number of Failures Before s Successes	Cumulative Probability	
8	0	=NEGBINOM.DIST($A8,$B$2,$B$3,1)	
9	1	=NEGBINOM.DIST($A9,$B$2,$B$3,1)	
10	2	=NEGBINOM.DIST($A10,$B$2,$B$3,1)	
11	3	=NEGBINOM.DIST($A11,$B$2,$B$3,1)	
12	4	=NEGBINOM.DIST($A12,$B$2,$B$3,1)	
13	5	=NEGBINOM.DIST($A13,$B$2,$B$3,1)	
14	6	=NEGBINOM.DIST($A14,$B$2,$B$3,1)	
15	7	=NEGBINOM.DIST($A15,$B$2,$B$3,1)	
16	8	=NEGBINOM.DIST($A16,$B$2,$B$3,1)	
17	9	=NEGBINOM.DIST($A17,$B$2,$B$3,1)	
18	10	=NEGBINOM.DIST($A18,$B$2,$B$3,1)	

	A	B	C
1	**Negative Binomial Distribution Parameters**		
2	Required Number of Successes (s)	5	
3	Probability of Success (p)	0.50	
4			Random Number
5	Randomly Generated Negative Binomial Value	6	0.6734
6			
7	Number of Failures Before s Successes	Cumulative Probability	
8	0	0.031	
9	1	0.109	
10	2	0.227	
11	3	0.363	
12	4	0.500	
13	5	0.623	
14	6	0.726	
15	7	0.806	
16	8	0.867	
17	9	0.910	
18	10	0.941	

MODEL *file*

negativebinomial

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 13.36 Excel Template to Generate Values from a Poisson Distribution

	A	B	C
1	**Poisson Distribution Parameter**		
2	Mean (m)	5	
3			Random Number
4	Randomly Generated Poisson Value	=XLOOKUP(C4,B7:B507,A7:A507,"error",1,2)	=RAND()
5			
6	Number of Event Occurrences	Cumulative Probability	
7	0	=POISSON.DIST($A7,$B$2,1)	
8	1	=POISSON.DIST($A8,$B$2,1)	
9	2	=POISSON.DIST($A9,$B$2,1)	
10	3	=POISSON.DIST($A10,$B$2,1)	
11	4	=POISSON.DIST($A11,$B$2,1)	
12	5	=POISSON.DIST($A12,$B$2,1)	
13	6	=POISSON.DIST($A13,$B$2,1)	
14	7	=POISSON.DIST($A14,$B$2,1)	
15	8	=POISSON.DIST($A15,$B$2,1)	
16	9	=POISSON.DIST($A16,$B$2,1)	
17	10	=POISSON.DIST($A17,$B$2,1)	

	A	B	C
1	**Poisson Distribution Parameter**		
2	Mean (m)	5	
3			Random Number
4	Randomly Generated Poisson Value	6	0.7038
5			
6	Number of Event Occurrences	Cumulative Probability	
7	0	0.007	
8	1	0.040	
9	2	0.125	
10	3	0.265	
11	4	0.440	
12	5	0.616	
13	6	0.762	
14	7	0.867	
15	8	0.932	
16	9	0.968	
17	10	0.986	

MODEL *file*

poisson

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 14

Linear Optimization Models

Contents

Analytics in Action: Stitch Fix

14.1 A Simple Maximization Problem
Problem Formulation
Mathematical Model for the Par, Inc. Problem

14.2 Solving the Par, Inc. Problem
The Geometry of the Par, Inc. Problem
Solving Linear Programs with Excel Solver

14.3 A Simple Minimization Problem
Problem Formulation
Solution for the M&D Chemicals Problem

14.4 Special Cases of Linear Program Outcomes
Alternative Optimal Solutions
Infeasibility
Unbounded

14.5 Sensitivity Analysis
Interpreting Excel Solver Sensitivity Report

14.6 General Linear Programming Notation and More Examples
Investment Portfolio Selection
Transportation Planning
Maximizing Banner Ad Revenue
Assigning Project Leaders to Clients
Diet Planning

14.7 Generating an Alternative Optimal Solution for a Linear Program

Summary 783
Glossary 784
Problems 785

Available in the Cengage eBook:
Appendix: Linear Programming with R

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Create a linear programming model based on a problem description.

LO 2 Create a spreadsheet model of a linear program and solve it with Excel Solver.

LO 3 Interpret the solution of a linear program.

LO 4 Calculate the value of the slack or surplus of a constraint.

LO 5 Perform sensitivity analysis on an objective function coefficient and interpret an objective-function-coefficient range.

LO 6 Perform sensitivity analysis on a right-hand-side of a constraint and interpret a right-hand-side range.

LO 7 Interpret a shadow price for a constraint.

LO 8 Find an alternative optimal solution to a linear program when one exists.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

Stitch Fix*

Stitch Fix is an online personalized styling service and apparel store founded in 2011. In 2021, Stitch Fix had revenue of $2.1 billion and over four million clients with operations in the United States and the United Kingdom. Stitch Fix clients fill out a style quiz, and are then paired with a personal stylist. Based on answers to the style quiz and direct feedback from the client, the stylist recommends items from over 1000 brands and styles. Stitch Fix clients can then select one of two ways to shop: "Fix" or "Freestyle." In the Fix option, the stylists select five items, known as a Fix, which can be previewed by the client online. The Fix is then sent to the client and the client can purchase or return any of the five items. In the Freestyle option, items are curated for the client by the stylist and placed in a personalized online store. The client can then shop in their personalized store. Regardless of the shopping option chosen by the client, success with the client is highly dependent on whether the stylist provides brands and styles that the client finds attractive and desirable enough to purchase. As such, making sure that the picks made by the stylist match the client's preferences is critical to the success of Stitch Fix.

Stitch Fix's website states "Our unique business model combines the human touch of expert stylists with the precision of advanced data science." Stitch Fix uses a system based on a linear programing model known as the assignment problem to assign stylists to clients to help ensure their success. The objective of the linear programming model is to maximize the expected number of happy clients. Based on millions of client-stylist interactions, data scientists have developed machine learning algorithms that estimate the probability that a client will be happy with each of its thousands of stylists. The linear programming model factors in the number of clients needing a stylist and the availability and capacity of the stylists over a given time period to assign stylists to clients to maximize expected positive client sentiment.

* Based on Schecter, J., "Add Constrained Optimization to Your Toolbelt," *stitchfix.com*.

This chapter begins our discussion of *prescriptive analytics* and how optimization models can be used to support and improve managerial decision making. Optimization problems maximize or minimize some function, called the **objective function**, and usually have a set of restrictions known as **constraints**. Consider the following typical applications of optimization:

1. A manufacturer wants to develop a production schedule and an inventory policy that will satisfy demand in future periods. Ideally, the schedule and policy will enable the company to satisfy demand and at the same time *minimize* the total production and inventory costs.
2. A financial analyst must select an investment portfolio from a variety of stock and bond investment alternatives. The analyst would like to establish the portfolio that *maximizes* the return on investment.
3. A marketing manager wants to determine how best to allocate a fixed advertising budget among alternative advertising media such as online, radio, television, and magazine. The manager would like to determine the media mix that *maximizes* advertising effectiveness.
4. A company has warehouses in a number of locations. Given specific customer demands, the company would like to determine how much each warehouse should ship to each customer so that total transportation costs are *minimized.*
5. A hospital needs to determine the work schedule of emergency room nurses for the next month. The hospital would like to *minimize* the amount of overtime it must pay to nurses while ensuring that the emergency room is fully staffed while also accommodating time-off requests and work rules for nurses.
6. An airline company needs to assign planes arriving at a particular airport to arrival gates. The airline company wants to *minimize* the total fuel required for planes to taxi from landing to their gates. The airline must also satisfy constraints that each gate can only have one arriving plane at a time and some gates can only accept certain sizes of arriving planes.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Each of these examples has a clear objective. The manufacturer wants to minimize costs; the financial analyst wants to maximize return on investment; the marketing manager wants to maximize advertising effectiveness; the company in example 4 wants to minimize total transportation costs; the hospital wants to minimize the amount of overtime it must pay nurses; and the airline company wants to minimize the cost of fuel required for taxiing planes.

Linear programming was initially referred to as "programming in a linear structure." In 1948, Tjalling Koopmans suggested to George Dantzig that the name was much too long: Koopman's suggestion was to shorten it to linear programming. George Dantzig agreed, and the field we now know as linear programming was named.

Likewise, each problem has constraints that limit the degree to which the objective can be pursued. In example 1, the manufacturer is restricted by the constraints requiring product demand to be satisfied and limiting production capacity. The financial analyst's portfolio problem is constrained by the total amount of investment funds available and the maximum amounts that can be invested in each stock or bond. The marketing manager's media selection decision is constrained by a fixed advertising budget and the availability of the various media. In the transportation problem, the minimum-cost shipping schedule is constrained by the supply of product available at each warehouse. For the hospital, the possible schedules for nurses are constrained by how many nurses are available as well as by time-off requests and existing work rules for nurses. For the airline company, each arriving plane must have an arrival gate, arrival gates can only accept one arriving plane at a time, and some gates cannot accept certain arriving planes due to their size.

Optimization models can be linear or nonlinear. We begin with linear optimization models, also known as linear programs. Linear programming is a problem-solving approach developed to help managers make better decisions. Numerous applications of linear programming can be found in today's competitive business environment. For instance, General Electric uses linear programming to help determine how much to invest in a variety of new solar power technologies, Stitch Fix uses linear programming to assign stylists to clients, and Marathon Oil Company uses linear programming for gasoline blending and to evaluate the economics of a new terminal or pipeline.

In this chapter, we also discuss how to use Excel Solver to solve linear programs. The online appendix that accompanies this chapter introduces the use of R to solve linear programs.

14.1 A Simple Maximization Problem

Par, Inc. is a small manufacturer of golf equipment and supplies whose management has decided to move into the market for medium- and high-priced golf bags. Par's distributor is enthusiastic about the new product line and has agreed to buy all the golf bags Par produces over the next three months.

After a thorough investigation of the steps involved in manufacturing a golf bag, management determined that each golf bag produced will require the following operations:

1. Cutting and dyeing the material
2. Sewing
3. Finishing (inserting umbrella holder, club separators, etc.)
4. Inspection and packaging

The director of manufacturing analyzed each of the operations and concluded that if the company produces a medium-priced standard model, each bag will require $\frac{7}{10}$ hour in the cutting and dyeing department, $\frac{1}{2}$ hour in the sewing department, 1 hour in the finishing department, and $\frac{1}{10}$ hour in the inspection and packaging department. The more expensive deluxe model will require 1 hour for cutting and dyeing, $\frac{5}{6}$ hour for sewing, $\frac{2}{3}$ hour for finishing, and $\frac{1}{4}$ hour for inspection and packaging. This production information is summarized in Table 14.1.

Par's production is constrained by a limited number of hours available in each department. After studying departmental workload projections, the director of manufacturing estimates that 630 hours for cutting and dyeing, 600 hours for sewing, 708 hours for finishing, and 135 hours for inspection and packaging will be available for the production of golf bags during the next three months.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 14.1 Production Requirements per Golf Bag

	Production Time (hours)	
Department	**Standard Bag**	**Deluxe Bag**
Cutting and Dyeing	7/10	1
Sewing	1/2	5/6
Finishing	1	2/3
Inspection and Packaging	1/10	1/4

It is important to understand that we are maximizing profit contribution, not profit. Overhead and other shared costs must be deducted before arriving at a profit figure.

The accounting department analyzed the production data, assigned all relevant variable costs, and arrived at prices for both bags that will result in a profit contribution[1] of $10 for every standard bag and $9 for every deluxe bag produced. Let us now develop a mathematical model of the Par, Inc. problem that can be used to determine the number of standard bags and the number of deluxe bags to produce in order to maximize total profit contribution.

Problem Formulation

Problem formulation, or **modeling**, is the process of translating the verbal statement of a problem into a mathematical statement. Formulating models is an art that can be mastered only with practice and experience. Even though every problem has some unique features, most problems also have common features. As a result, *some* general guidelines for optimization model formulation can be helpful, especially for beginners. We will illustrate these general guidelines by developing a mathematical model for Par, Inc.

Understand the problem thoroughly We selected the Par, Inc. problem to introduce linear programming because it is easy to understand. However, more complex problems will require much more effort to identify the items that need to be included in the model. In such cases, read the problem description to get a feel for what is involved. Taking notes will help you focus on the key issues and facts.

Describe the objective The objective is to maximize the total contribution to profit.

Describe each constraint Four constraints relate to the number of hours of manufacturing time available; they restrict the number of standard bags and the number of deluxe bags that can be produced.

- *Constraint 1:* The number of hours of cutting and dyeing time used must be less than or equal to the number of hours of cutting and dyeing time available.
- *Constraint 2:* The number of hours of sewing time used must be less than or equal to the number of hours of sewing time available.
- *Constraint 3:* The number of hours of finishing time used must be less than or equal to the number of hours of finishing time available.
- *Constraint 4:* The number of hours of inspection and packaging time used must be less than or equal to the number of hours of inspection and packaging time available.

Define the decision variables The controllable inputs for Par, Inc. are (1) the number of standard bags produced and (2) the number of deluxe bags produced. Let:

$$S = \text{number of standard bags}$$
$$D = \text{number of deluxe bags}$$

In optimization terminology, S and D are referred to as the **decision variables**.

[1]From an accounting perspective, profit contribution is more correctly described as the contribution margin per bag since overhead and other shared costs are not allocated.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Write the objective in terms of the decision variables Par's profit contribution comes from two sources: (1) the profit contribution made by producing S standard bags and (2) the profit contribution made by producing D deluxe bags. If Par makes \$10 for every standard bag, the company will make \$10S if S standard bags are produced. Also, if Par makes \$9 for every deluxe bag, the company will make \$9D if D deluxe bags are produced. Thus, we have

$$\text{Total profit contribution} = 10S + 9D$$

Because the objective—maximize total profit contribution—is a function of the decision variables S and D, we refer to $10S + 9D$ as the *objective function.* Using *Max* as an abbreviation for maximize, we write Par's objective as follows:

$$\text{Max } 10S + 9D$$

The units of measurement on the left-hand side of the constraint must match the units of measurement on the right-hand side.

Write the constraints in terms of the decision variables
Constraint 1:

$$\begin{pmatrix}\text{Hours of cutting and}\\ \text{dyeing time used}\end{pmatrix} \leq \begin{pmatrix}\text{Hours of cutting and}\\ \text{dyeing time available}\end{pmatrix}$$

Every standard bag Par produces will use $\frac{7}{10}$ hour cutting and dyeing time; therefore, the total number of hours of cutting and dyeing time used in the manufacture of S standard bags is $\frac{7}{10}S$. In addition, because every deluxe bag produced uses 1 hour of cutting and dyeing time, the production of D deluxe bags will use $1D$ hours of cutting and dyeing time. Thus, the total cutting and dyeing time required for the production of S standard bags and D deluxe bags is given by

$$\text{Total hours of cutting and dyeing time used} = \tfrac{7}{10}S + 1D$$

The director of manufacturing stated that Par has at most 630 hours of cutting and dyeing time available. Therefore, the production combination we select must satisfy the requirement:

$$\tfrac{7}{10}S + 1D \leq 630 \qquad \textbf{(14.1)}$$

Constraint 2:

$$\begin{pmatrix}\text{Hours of sewing}\\ \text{time used}\end{pmatrix} \leq \begin{pmatrix}\text{Hours of sewing}\\ \text{time available}\end{pmatrix}$$

From Table 14.1, we see that every standard bag manufactured will require $\frac{1}{2}$ hour for sewing, and every deluxe bag will require $\frac{5}{6}$ hour for sewing. Because 600 hours of sewing time are available, it follows that

$$\tfrac{1}{2}S + \tfrac{5}{6}D \leq 600 \qquad \textbf{(14.2)}$$

Constraint 3:

$$\begin{pmatrix}\text{Hours of finishing}\\ \text{time used}\end{pmatrix} \leq \begin{pmatrix}\text{Hours of finishing}\\ \text{time available}\end{pmatrix}$$

Every standard bag manufactured will require 1 hour for finishing, and every deluxe bag will require $\frac{2}{3}$ hour for finishing. With 708 hours of finishing time available, it follows that

$$1S + \tfrac{2}{3}D \leq 708 \qquad \textbf{(14.3)}$$

Constraint 4:

$$\begin{pmatrix}\text{Hours of inspection and}\\ \text{packaging time used}\end{pmatrix} \leq \begin{pmatrix}\text{Hours of inspection and}\\ \text{packaging time available}\end{pmatrix}$$

Every standard bag manufactured will require $\frac{1}{10}$ hour for inspection and packaging, and every deluxe bag will require $\frac{1}{4}$ hour for inspection and packaging. Because 135 hours of inspection and packaging time are available, it follows that

$$\tfrac{1}{10}S + \tfrac{1}{4}D \leq 135 \qquad \textbf{(14.4)}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

We have now specified the mathematical relationships for the constraints associated with the four departments. However, we must also remember that clearly, Par cannot produce a negative number of standard or deluxe bags. Thus, to prevent the decision variables S and D from having negative values, two constraints must be added:

$$S \geq 0 \quad \text{and} \quad D \geq 0 \tag{14.5}$$

These constraints ensure that the solution to the problem will contain only nonnegative values for the decision variables and are thus referred to as the **nonnegativity constraints**. Nonnegativity constraints are a general feature of many linear programming problems and may be written in the abbreviated form:

$$S, D \geq 0$$

Mathematical Model for the Par, Inc. Problem

The mathematical statement, or mathematical formulation, of the Par, Inc. problem is now complete. We succeeded in translating the objective and constraints of the problem into a set of mathematical relationships, referred to as a **mathematical model**. The complete mathematical model for the Par, Inc. problem is as follows:

$$\begin{aligned}
&\text{Max} && 10S + 9D && \\
&\text{subject to (s.t.)} && && \\
&&& \tfrac{7}{10}S + 1D \leq 630 && \text{Cutting and dyeing} \\
&&& \tfrac{1}{2}S + \tfrac{5}{6}D \leq 600 && \text{Sewing} \\
&&& 1S + \tfrac{2}{3}D \leq 708 && \text{Finishing} \\
&&& \tfrac{1}{10}S + \tfrac{1}{4}D \leq 135 && \text{Inspection and packaging} \\
&&& S, D \geq 0 &&
\end{aligned}$$

Our job now is to find the product mix (i.e., the combination of values for S and D) that satisfies all the constraints and at the same time yields a value for the objective function that is greater than or equal to the value given by any other feasible solution. Once these values are calculated, we will have found the optimal solution to the problem.

This mathematical model of the Par, Inc. problem is a **linear programming model**, or **linear program**, because the objective function and all constraint functions (the left-hand sides of the constraint inequalities) are linear functions of the decision variables.

Mathematical functions in which each variable appears in a separate term and is raised to the first power are called **linear functions**. The objective function ($10S + 9D$) is linear because each decision variable appears in a separate term and has an exponent of 1. The amount of production time required in the cutting and dyeing department ($\frac{7}{10}S + 1D$) is also a linear function of the decision variables for the same reason. Similarly, the functions on the left-hand side of all the constraint inequalities (the constraint functions) are linear functions. Thus, the mathematical formulation of this problem is referred to as a linear program.

Notes + Comments

1. The three assumptions necessary for a linear programming model to be appropriate are proportionality, additivity, and divisibility. *Proportionality* means that the contribution to the objective function and the amount of resources used in each constraint are proportional to the value of each decision variable. *Additivity* means that the value of the objective function and the total resources used can be found by summing the objective function contribution and the resources used for all decision variables. *Divisibility* means that the decision variables are continuous. The divisibility assumption plus the nonnegativity constraints mean that decision variables can take on any value greater than or equal to zero.
2. Linear programming has nothing to do with computer programming. The use of the word "programming" means "choosing a course of action." Linear programming involves choosing a course of action when the mathematical model of the problem contains only linear functions.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

14.2 Solving the Par, Inc. Problem

Now that we have modeled the Par, Inc. problem as a linear program, let us discuss how we might find the optimal solution. The optimal solution must be a feasible solution. A **feasible solution** is a setting of the decision variables that satisfies all of the constraints of the problem. The optimal solution also must have an objective function value as good as any other feasible solution. For a maximization problem like Par, Inc., this means that the solution must be feasible and achieve the highest objective function value of any feasible solution. To solve a linear program then, we must search over the **feasible region**, which is the set of all feasible solutions, and find the solution that gives the best objective function value.

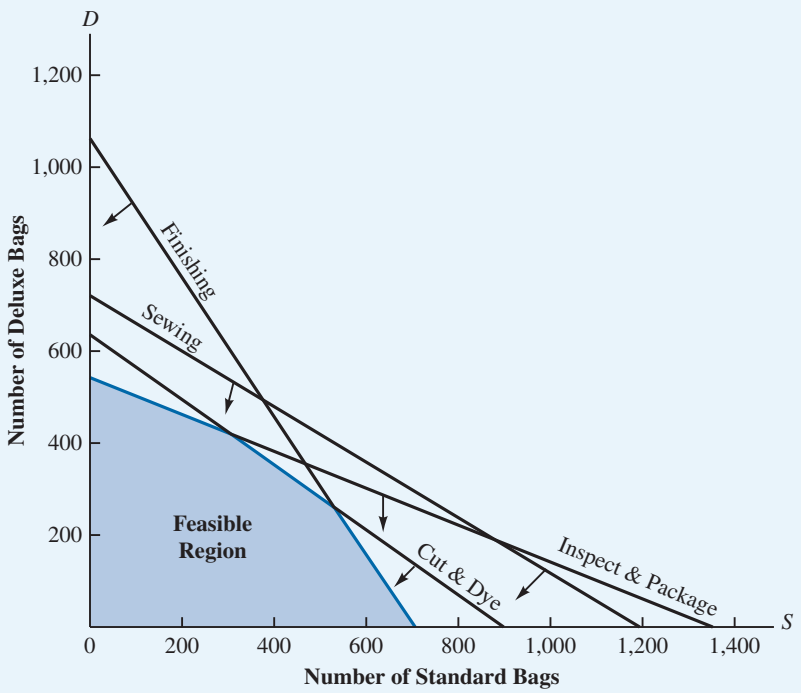
Because the Par, Inc. model has two decision variables, we are able to graph the feasible region. Discussing the geometry of the feasible region of the model will help us better understand linear programming and how we are able to solve much larger problems on the computer.

The Geometry of the Par, Inc. Problem

Recall that the feasible region is the set of points that satisfies all of the constraints of the problem. When we have only two decision variables and the functions of these variables are linear, they form lines in two-dimensional space. If the constraints are inequalities, the constraint cuts the space into two, with the line and the area on one side of the line being the space that satisfies that constraint. These subregions are called *half spaces*. The *intersection* of these half spaces makes up the feasible region.

The feasible region for the Par, Inc. problem is shown in Figure 14.1. Notice that the horizontal axis corresponds to the value of S and the vertical axis to the value of D. The nonnegativity constraints define that the feasible region is in the area bounded by the horizontal and vertical axes. Each of the four constraints is graphed as equality (a line), and arrows show the direction of the half space that satisfies the inequality constraint. The intersection of the four half spaces in the area bounded by the axes is the shaded region; this is the feasible region for the Par, Inc. problem. Any point in the shaded region satisfies all four constraints of the problem and nonnegativity.

Figure 14.1 Feasible Region for the Par, Inc. Problem



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To solve the Par, Inc. problem, we must find the point in the feasible region that results in the highest possible objective function value. A contour line is a set of points on a map, all of which have the same elevation. Similar to the way contour lines are used in geography, we may define an *objective function contour* to be a set of points (in this case a line) that yield a fixed value of the objective function. By choosing a fixed value of the objective function, we may plot contour lines of the objective function over the feasible region (Figure 14.2). In this case, as we move away from the origin we see higher values of the objective function and the highest such contour is $10S + 9D = 7{,}668$, after which we leave the feasible region. The highest value contour intersects the feasible region at a single point—point ③.

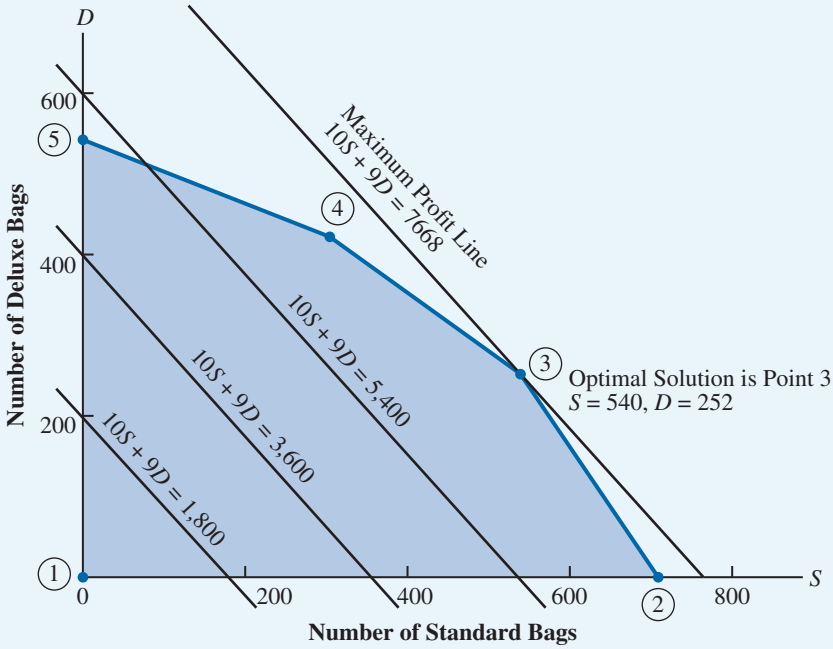
Of course, this geometric approach to solving a linear program is limited to problems with only two variables. What have we learned that can help us solve larger linear optimization problems?

Based on the geometry of Figure 14.2, to solve a linear optimization problem we only have to search over the **extreme points** of the feasible region to find an optimal solution. The extreme points are found where constraints intersect on the boundary of the feasible region. In Figure 14.2, points ①, ②, ③, ④, and ⑤ are the extreme points of the feasible region.

Because each extreme point lies at the intersection of two constraint lines, we may obtain the values of S and D by solving simultaneously as equalities, the pair of constraints that form the given point. The values of S and D and the objective function value at points ① through ⑤ are as follows:

Point	S	D	Profit $= 10S + 9D$
1	0	0	10(0) + 9(0) = 0
2	708	0	10(708) + 9(0) = 7,080
3	540	252	10(540) + 9(252) = 7,668
4	300	420	10(300) + 9(420) = 6,780
5	0	540	10(0) + 9(540) = 4,860

Figure 14.2 The Optimal Solution to the Par, Inc. Problem



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The highest profit is achieved at point ③. Therefore, the optimal plan is to produce 540 standard bags and 252 deluxe bags, as shown in Figure 14.2.

It turns out that this approach of investigating only extreme points works well and generalizes for larger problems. The simplex algorithm, developed by George Dantzig, is quite effective at investigating extreme points in an intelligent way to find the optimal solution to even very large linear programs.

Excel Solver utilizes Dantzig's simplex algorithm to solve linear programs by systematically finding which set of constraints form the optimal extreme point of the feasible region. Once it finds an optimal solution, Solver then reports the optimal values of the decision variables and the optimal objective function value. Let us illustrate now how to use Excel Solver to find the optimal solution to the Par, Inc. problem.

Solving Linear Programs with Excel Solver

The first step in solving a linear optimization model in Excel is to construct the relevant what-if model. Using the principles for developing good spreadsheet models discussed in Chapter 12, a what-if model for optimization allows the user to try different values of the decision variables and see easily (a) whether that trial solution is feasible and (b) the value of the objective function for that trial solution.

Figure 14.3 shows a spreadsheet model for the Par, Inc. problem with a trial solution of one standard bag and one deluxe bag. Rows 1 through 10 contain the parameters for the problem. Row 14 contains the decision variable cells: Cells B14 and C14 are the locations for the number of standard and deluxe bags to produce. Cell B16 calculates the objective function value for the trial solution by using the SUMPRODUCT function. The SUMPRODUCT function is very useful for linear problems. Recall how the SUMPRODUCT function works:

$$=\text{SUMPRODUCT(B9:C9,\$B\$14:\$C\$14)} = B9 * B14 + C9 * C14 = 10(1) + 9(1) = 19$$

We likewise use the SUMPRODUCT function in cells B19:B22 to calculate the number of hours used in each of the four departments. The hours available are immediately to the right for each department. Hence, we see that the current solution is feasible, since Hours Used do not exceed Hours Available in any department.

Once the what-if model is built, we need a way to convey to Excel Solver the structure of the linear optimization model. This is accomplished through the Excel Solver dialog box as follows:

Step 1. Click the **Data** tab in the Ribbon
Step 2. Click **Solver** in the **Analyze** group
Step 3. When the **Solver Parameters** dialog box appears (Figure 14.4):
Enter *B16* in the **Set Objective:** box
Select **Max** for the **To:** option
Enter *B14:C14* in the **By Changing Variable Cells:** box
Step 4. Click the **Add** button
When the **Add Constraint** dialog box appears:
Enter *B19:B22* in the left-hand box under **Cell Reference:**
Select **<=** from the drop-down button
Enter *C19:C22* in the **Constraint**: box
Click **OK**
Step 5. Select the checkbox for **Make Unconstrained Variables Non-Negative**
Step 6. From the drop-down menu for **Select a Solving Method:**, choose **Simplex LP**
Step 7. Click **Solve**
Step 8. When the **Solver Results** dialog box appears:
Select **Keep Solver Solution**
In the **Reports** section, select **Answer Report**
Click **OK**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.3 What-If Spreadsheet Model for Par, Inc.

	A	B	C	D	E
1	**Par, Inc.**				
2	**Parameters**				
3		**Production Time (Hours)**		**Time Available**	
4	**Operation**	**Standard**	**Deluxe**	**Hours**	
5	Cutting and Dyeing	=7/10	1	630	
6	Sewing	=5/10	=5/6	600	
7	Finishing	1	=2/3	708	
8	Inspection and Packaging	=1/10	=1/4	135	
9	Profit Per Bag	10	9		
10					
11	**Model**				
12					
13		**Standard**	**Deluxe**		
14	Bags Produced	1	1		
15					
16	**Total Profit**	=SUMPRODUCT(B9:C9,$B$14:$C$14)			
17					
18	**Department**	**Hours Used**	**Hours Available**		
19	Cutting and Dyeing	=SUMPRODUCT($B$14:$C$14,B5:C5)	=D5		
20	Sewing	=SUMPRODUCT($B$14:$C$14,B6:C6)	=D6		
21	Finishing	=SUMPRODUCT($B$14:$C$14,B7:C7)	=D7		
22	Inspection and Packaging	=SUMPRODUCT($B$14:$C$14,B8:C8)	=D8		
23					

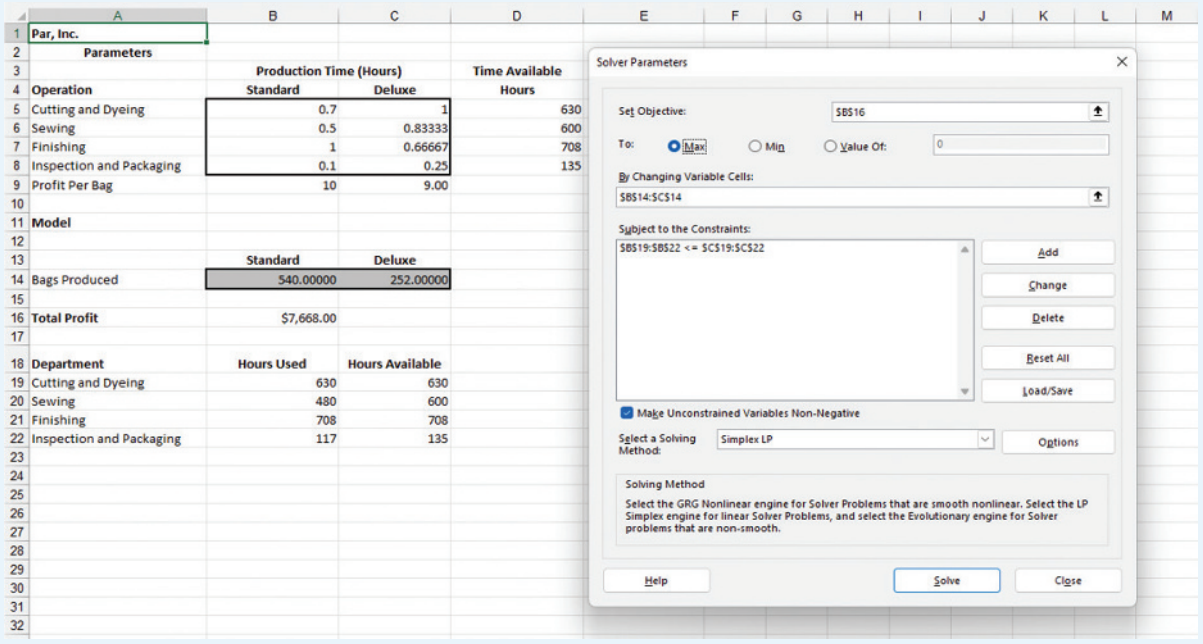
	A	B	C	D	E
1	**Par, Inc.**				
2	**Parameters**				
3		**Production Time (Hours)**		**Time Available**	
4	**Operation**	**Standard**	**Deluxe**	**Hours**	
5	Cutting and Dyeing	0.7	1	630	
6	Sewing	0.5	0.83333	600	
7	Finishing	1	0.66667	708	
8	Inspection and Packaging	0.1	0.25	135	
9	Profit Per Bag	10	9.00		
10					
11	**Model**				
12					
13		**Standard**	**Deluxe**		
14	Bags Produced	1.00	1.00		
15					
16	**Total Profit**	$19.00			
17					
18	**Department**	**Hours Used**	**Hours Available**		
19	Cutting and Dyeing	1.7	630		
20	Sewing	1.333333333	600		
21	Finishing	1.666666667	708		
22	Inspection and Packaging	0.35	135		
23					

MODEL *file*

par

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.4 Solver Dialog Box and Solution to the Par, Inc. Problem



	A	B	C	D
1	Par, Inc.			
2	Parameters			
3		Production Time (Hours)		Time Available
4	Operation	Standard	Deluxe	Hours
5	Cutting and Dyeing	0.7	1	630
6	Sewing	0.5	0.83333	600
7	Finishing	1	0.66667	708
8	Inspection and Packaging	0.1	0.25	135
9	Profit Per Bag	10	9.00	
10				
11	Model			
12				
13		Standard	Deluxe	
14	Bags Produced	540.00000	252.00000	
15				
16	Total Profit	$7,668.00		
17				
18	Department	Hours Used	Hours Available	
19	Cutting and Dyeing	630	630	
20	Sewing	480	600	
21	Finishing	708	708	
22	Inspection and Packaging	117	135	

The completed Solver dialog box and solution for the Par, Inc. problem are shown in Figure 14.4. The optimal solution is to make 540 standard bags and 252 deluxe bags (see cells B14 and C14) for a profit of $7,688 (see cell B16). This corresponds to point ③ in Figure 14.2. Also note that, from cells B19:B22 compared to C19:C22, we use all cutting and dyeing time as well as all finishing time. This is, of course, consistent with what we have seen in Figures 14.1 and 14.2: The cutting, dyeing, and finishing constraints intersect to form point ③ in the graph.

The Excel Solver Answer Report appears in Figure 14.5. The Answer Report contains three sections: Objective Cell, Variable Cells, and Constraints. In addition to some other information, each section gives the cell location, name, and value of the cell(s). The Objective Cell section indicates that the optimal (Final Value) of Total Profit is $7,668.00. In the Variable Cells section, the two far-right columns indicate the optimal values of the decision cells and whether or not the variables are required to be integer (here they are labeled "Contin" for continuous). Note that Solver generates a Name for a cell by concatenating the text to the left and above that cell. Hence, the name of cell $B$14 is created by combining the labels "Bags Produced" and "Standard" to produce the name "Bags Produced Standard."

The Constraints section gives the left-hand side value for each constraint (in this case the hours used), the formula showing the constraint relationship, the status (Binding or Not Binding), and the Slack value. A **binding constraint** is one that holds as an equality at the optimal solution. Geometrically, binding constraints intersect to form the optimal point. We see in Figure 14.5 that the cutting and dyeing and finishing constraints are designated as binding, consistent with our geometric study of this problem.

The **slack** value for each less-than-or-equal-to constraint indicates the difference between the left-hand and right-hand values for a constraint. Of course, by definition,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.5 The Solver Answer Report for the Par, Inc. Problem

	A	B	C	D	E	F	G
13							
14	Objective Cell (Max)						
15		Cell	Name	Original Value	Final Value		
16		$B$16	Total Profit	$19.00	$7,668.00		
17							
18							
19	Variable Cells						
20		Cell	Name	Original Value	Final Value	Integer	
21		$B$14	Bags Produced Standard	1.00000	540.00000	Contin	
22		$C$14	Bags Produced Deluxe	1.00000	252.00000	Contin	
23							
24							
25	Constraints						
26		Cell	Name	Cell Value	Formula	Status	Slack
27		$B$19	Cutting and Dyeing Hours Used	630	$B$19<=$C$19	Binding	0
28		$B$20	Sewing Hours Used	480	$B$20<=$C$20	Not Binding	120
29		$B$21	Finishing Hours Used	708	$B$21<=$C$21	Binding	0
30		$B$22	Inspection and Packaging Hours Used	117	$B$22<=$C$22	Not Binding	18
31							

binding constraints have zero slack. Consider for example the sewing department constraint. By adding a nonnegative **slack variable**, we can make the constraint equality:

$$\tfrac{1}{2}S + \tfrac{5}{6}D \leq 600$$
$$\tfrac{1}{2}S + \tfrac{5}{6}D + \text{slack}_{\text{sewing}} = 600$$
$$\text{slack}_{\text{sewing}} = 600 - \tfrac{1}{2}(540) + \tfrac{5}{6}(252) = 600 - 270 - 210 = 120$$

The slack value for the inspecting and packaging constraint is calculated in a similar way. For resource constraints like departmental hours available, the slack value gives the amount of unused resource, in this case, time measured in hours.

Notes + Comments

1. Notice in the data section for the Par, Inc. spreadsheet, shown in Figure 14.3, that we have entered fractions in cells C6: *=5/6* and C7: *=2/3*. We do this to make sure we maintain accuracy because rounding these values could have an impact on our solution.
2. By selecting **Make Unconstrained Variables Non-Negative** in the **Solver Parameters** dialog box, all decision variables are constrained to be nonnegative.
3. Although we have shown the Answer Report and how to interpret it, we will usually show the solution to an optimization problem directly in the spreadsheet. A well-designed spreadsheet that follows the principles discussed in Chapter 12 should make it easy for the user to interpret the optimal solution directly from the spreadsheet.
4. In addition to the Answer Report, Solver also allows you to generate two other reports. The Sensitivity Report will be discussed in Section 14.5. The Limits Report gives information on the objective function value when variables are set to their limits.
5. In the Par, Inc. example, it was coincidental that the optimal value of each type of golf bag to produce was integer valued. If the optimal values of the decision variables were fractional, we would need to add further constraints requiring these values be integer. We discuss how to do this in Chapter 15.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

14.3 A Simple Minimization Problem

M&D Chemicals produces two products that are sold as raw materials to companies that manufacture bath soaps and laundry detergents. Based on an analysis of current inventory levels and potential demand for the coming month, M&D's management specified that the combined production for products A and B must total at least 350 gallons. Separately, a major customer's order for 125 gallons of product A must also be satisfied. Product A requires 2 hours of processing time per gallon, and product B requires 1 hour of processing time per gallon. For the coming month, 600 hours of processing time are available. M&D's objective is to satisfy these requirements at a minimum total production cost. Production costs are $2 per gallon for product A and $3 per gallon for product B.

Problem Formulation

To find the minimum-cost production schedule, we will formulate the M&D Chemicals problem as a linear program. Following a procedure similar to the one used for the Par, Inc. problem, we first define the decision variables and the objective function for the problem. Let

A = number of gallons of product A to produce

B = number of gallons of product B to produce

With production costs at $2 per gallon for product A and $3 per gallon for product B, the objective function that corresponds to the minimization of the total production cost can be written as

$$\text{Min } 2A + 3B$$

Next consider the constraints placed on the M&D Chemicals problem. To satisfy the major customer's demand for 125 gallons of product A, we know A must be at least 125. Thus, we write the constraint

$$1A \geq 125$$

For the combined production for both products, which must total at least 350 gallons, we can write the constraint

$$1A + 1B \geq 350$$

Finally, for the limitation of 600 hours on available processing time, we add the constraint

$$2A + 1B \leq 600$$

After adding the nonnegativity constraints ($A, B \geq 0$), we arrive at the following linear program for the M&D Chemicals problem:

$$\begin{aligned} \text{Min} \quad & 2A + 3B \\ \text{s.t.} \quad & \\ & 1A \geq 125 && \text{Demand for product A} \\ & 1A + 1B \geq 350 && \text{Total production} \\ & 2A + 1B \leq 600 && \text{Processing time} \\ & A, B \geq 0 \end{aligned}$$

Solution for the M&D Chemicals Problem

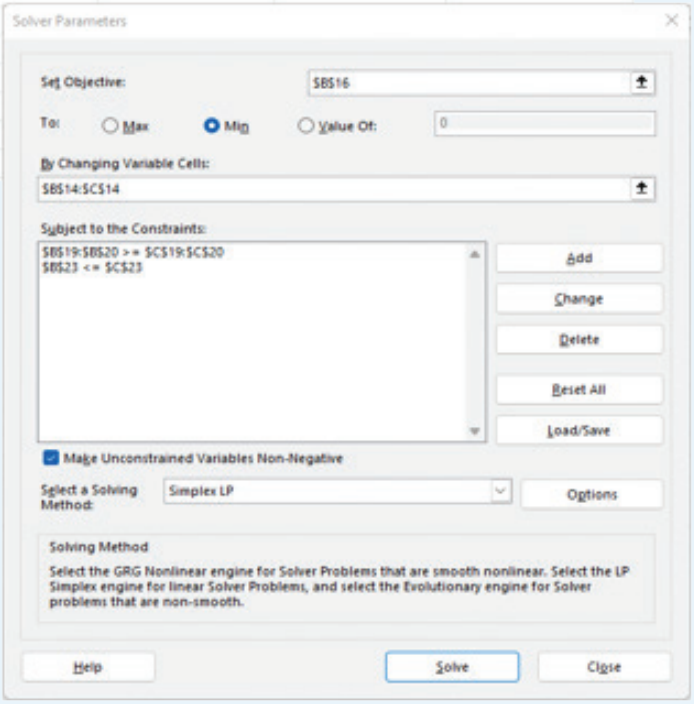
MODEL file
mdmodel

A spreadsheet model for the M&D Chemicals problem along with the Solver dialog box are shown in Figure 14.6. The complete linear programming model for the M&D Chemicals problem in Excel Solver is contained in the file *mdmodel*. We use the SUMPRODUCT function to calculate total cost in cell B16 and also to calculate total processing hours used in cell B23. The optimal solution, which is shown in the spreadsheet and in the Answer Report in Figure 14.7, is to make 250 gallons of product A and 100 gallons of product B,

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.6 Solver Dialog Box and Solution to the M&D Chemicals Problem

	A	B	C	D
1	**M&D Chemicals**			
2	**Parameters**			
3		Product A	Product B	Time Available
4	Processing Time (hours)	2	1	600
5	Production Cost	$2.00	$3.00	
6				
7	Minimum Total Production	350		
8	Product A Minimum	125		
9				
10				
11	**Model**			
12				
13		Product A	Product B	
14	Gallons Produced	250	100	
15				
16	Minimize Total Cost	$800.00		
17				
18		Provided	Required	
19	Product A	250	125	
20	Total Production	350	350	
21				
22		Hours Used	Hours Available	Unused Hours
23	Processing Time	600	600	0
24				
25				
26				
27				
28				
29				
30				



MODEL *file*

mdmodel

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.7 The Solver Answer Report for the M&D Chemicals Problem

Objective Cell (Min)

Cell	Name	Original Value	Final Value
$B$16	Minimize Total Cost	$0.00	$800.00

Variable Cells

Cell	Name	Original Value	Final Value	Integer
$B$14	Gallons Produced Product A	0	250	Contin
$C$14	Gallons Produced Product B	0	100	Contin

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$B$19	Product A Provided	250	$B$19>=$C$19	Not Binding	125
$B$20	Total Production Provided	350	$B$20>=$C$20	Binding	0
$B$23	Processing Time Hours Used	600	$B$23<=$C$23	Binding	0

for a total cost of $800. Both the total production constraint and the processing time constraints are binding (350 gallons are provided, the same as required, and all 600 processing hours are used). The requirement that at least 125 gallons of Product A be produced is not binding. For greater-than-or-equal-to constraints, we can define a nonnegative variable called a surplus variable. A **surplus variable** tells how much over the right-hand side the left-hand side of a greater-than-or-equal-to constraint is for a solution. A surplus variable is subtracted from the left-hand side of the constraint. For example,

$$1A \geq 125$$
$$1A - \text{surplus}_A = 125$$
$$\text{surplus}_A = 1A - 125 = 250 - 125 = 125$$

As was the case with less-than-or-equal-to constraints and slack variables, a positive value for a surplus variable indicates that the constraint is not binding.

Notes + Comments

1. In the spreadsheet and Solver model for the M&D Chemicals problem, we separated the greater-than-or-equal-to constraints and the less-than-or-equal-to constraints. This allows for easier entry of the constraints into the **Add Constraint** dialog box.
2. In the Excel Answer Report, both slack and surplus variables are labeled "Slack."
3. Like the Par, Inc. example, in the M&D Chemical example, the optimal values of the decision variables are integers. In the M&D Chemical example, fractional values of the decision variables might be acceptable as they represent quantities of liquid, so these variables can be modeled as continuous.

14.4 Special Cases of Linear Program Outcomes

In this section, we discuss three special situations that can arise when we attempt to solve linear programming problems.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Alternative Optimal Solutions

From the discussion of the graphical solution procedure, we know that optimal solutions can be found at the extreme points of the feasible region. Now let us consider the special case in which the optimal objective function contour line coincides with one of the binding constraint lines on the boundary of the feasible region. We will see that this situation can lead to the case of **alternative optimal solutions**; in such cases, more than one solution provides the optimal value for the objective function.

To illustrate the case of alternative optimal solutions, we return to the Par, Inc. problem. However, let us assume that the profit for the standard golf bag (S) has been decreased to \$6.30. The revised objective function becomes $6.3S + 9D$. The graphical solution of this problem is shown in Figure 14.8. Note that the optimal solution still occurs at an extreme point. In fact, it occurs at two extreme points: extreme point ④ ($S = 300$, $D = 420$) and extreme point ③ ($S = 540$, $D = 252$).

The objective function values at these two extreme points are identical; that is,

$$6.3S + 9D = 6.3(300) + 9(420) = 5{,}670$$

and

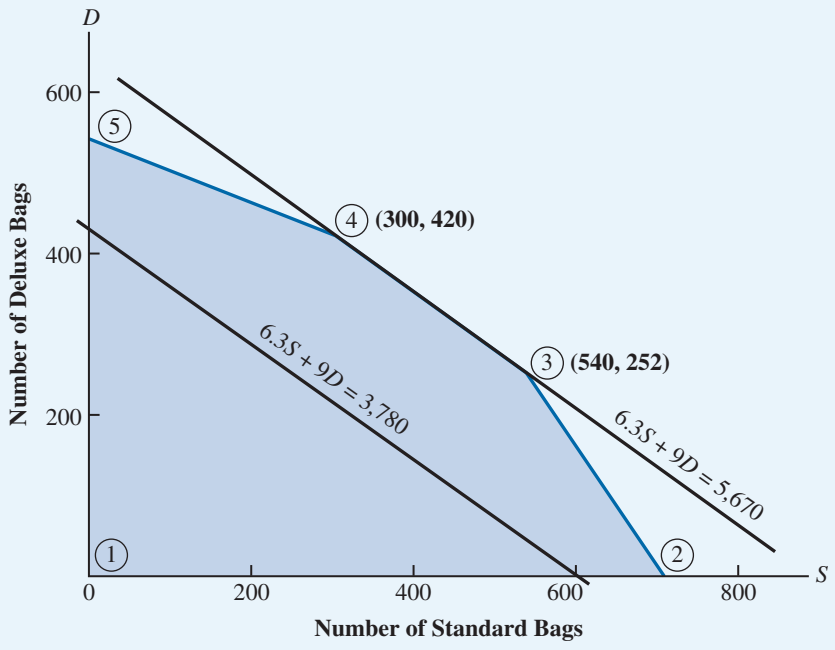
$$6.3S + 9D = 6.3(540) + 9(252) = 5{,}670$$

Furthermore, any point on the line connecting the two optimal extreme points also provides an optimal solution. For example, the solution point ($S = 420$, $D = 336$), which is halfway between the two extreme points, also provides the optimal objective function value of

$$6.3S + 9D = 6.3(420) + 9(336) = 5{,}670$$

A linear programming problem with alternative optimal solutions is generally a good situation for the manager or decision maker. It means that several combinations of

Figure 14.8 Par, Inc. Problem with an Objective Function of $6.3S + 9D$ (Alternative Optimal Solutions)



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the decision variables are optimal and that the manager can select the most desirable optimal solution. Unfortunately, determining whether a problem has alternative optimal solutions is not a simple matter. In Section 14.7, we discuss an approach for finding alternative optima.

Infeasibility

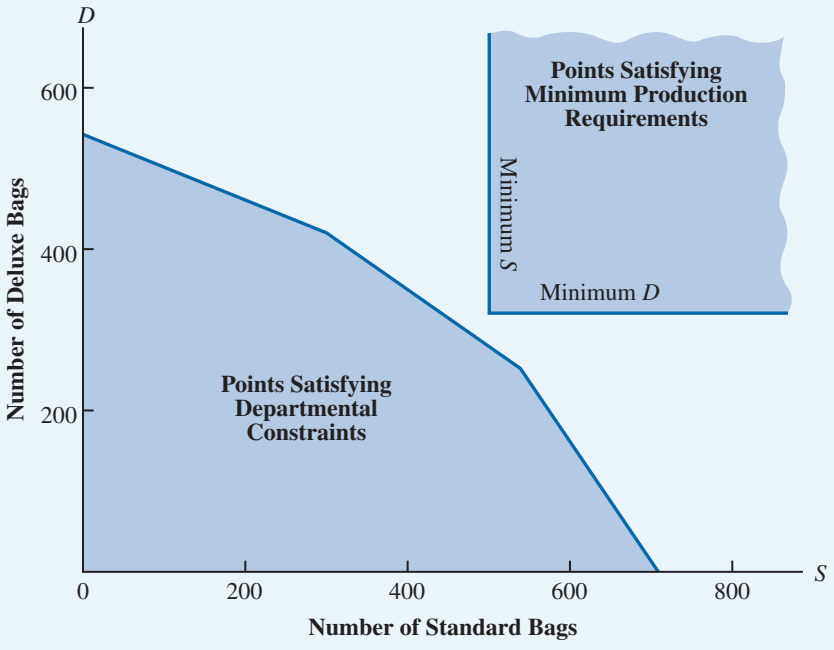
Problems with no feasible solution do arise in practice, most often because management's expectations are too high or because too many restrictions have been placed on the problem.

Infeasibility means that no solution to the linear programming problem satisfies all the constraints, including the nonnegativity conditions. Graphically, infeasibility means that a feasible region does not exist; that is, no points satisfy all the constraints and the nonnegativity conditions simultaneously. To illustrate this situation, let us look again at the problem faced by Par, Inc.

Suppose that management specified that at least 500 of the standard bags and at least 360 of the deluxe bags must be manufactured. The graph of the solution region may now be constructed to reflect these new requirements (see Figure 14.9). The shaded area in the lower left-hand portion of the graph depicts the points that satisfy the departmental constraints on the availability of time. The shaded area in the upper right-hand portion depicts the points that satisfy the minimum production requirements of 500 standard and 360 deluxe bags. But no points satisfy both sets of constraints. Thus, we see that if management imposes these minimum production requirements, no feasible region exists for the problem.

How should we interpret infeasibility in terms of this current problem? First, we should tell management that, given the resources available (i.e., production time for cutting and dyeing, sewing, finishing, and inspection and packaging), it is not possible to make 500 standard bags and 360 deluxe bags. Moreover, we can tell management exactly how much of each resource must be expended to make it possible to manufacture these numbers

Figure 14.9 No Feasible Region for the Par, Inc. Problem with Minimum Production Requirements of 500 Standard and 360 Deluxe Bags



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 14.2 Resources Needed to Manufacture 500 Standard Bags and 360 Deluxe Bags

Operation	Minimum Required Resources (hours)	Available Resources (hours)	Additional Resources Needed (hours)
Cutting and Dyeing	$\frac{7}{10}(500) + 1(360) = 710$	630	80
Sewing	$\frac{1}{2}(500) + \frac{5}{6}(360) = 550$	600	None
Finishing	$1(500) + \frac{2}{3}(360) = 740$	708	32
Inspection and Packaging	$\frac{1}{10}(500) + \frac{1}{4}(360) = 140$	135	5

of bags. Table 14.2 shows the minimum amounts of resources that must be available, the amounts currently available, and additional amounts that would be required to accomplish this level of production. Thus, we need 80 more hours for cutting and dyeing, 32 more hours for finishing, and 5 more hours for inspection and packaging to meet management's minimum production requirements.

If after reviewing this information, management still wants to manufacture 500 standard and 360 deluxe bags, additional resources must be provided. Perhaps the resource requirements can be met by hiring another person to work in the cutting and dyeing department, transferring a person from elsewhere in the plant to work part time in the finishing department, or having the sewing people help out periodically with the inspection and packaging. As you can see, once we discover the lack of a feasible solution, many possibilities are available for corrective management action. The important thing to realize is that linear programming analysis can help determine whether management's plans are feasible. By analyzing the problem using linear programming, we are often able to point out infeasible conditions and initiate corrective action.

Whenever you attempt to solve a problem that is infeasible, Excel Solver will return a message in the Solver Results dialog box, indicating that no feasible solution exists. In this case, you know that no solution to the linear programming problem will satisfy all constraints, including the nonnegativity conditions. Careful inspection of your formulation is necessary to try to identify why the problem is infeasible. In some situations, the only reasonable approach is to drop one or more constraints and re-solve the problem. If you are able to find an optimal solution for this revised problem, you will know that the constraint(s) that were omitted, in conjunction with the others, are causing the problem to be infeasible.

Unbounded

The solution to a maximization linear programming problem is **unbounded** if the value of the solution may be made infinitely large without violating any of the constraints; for a minimization problem, the solution is unbounded if the value may be made infinitely small.

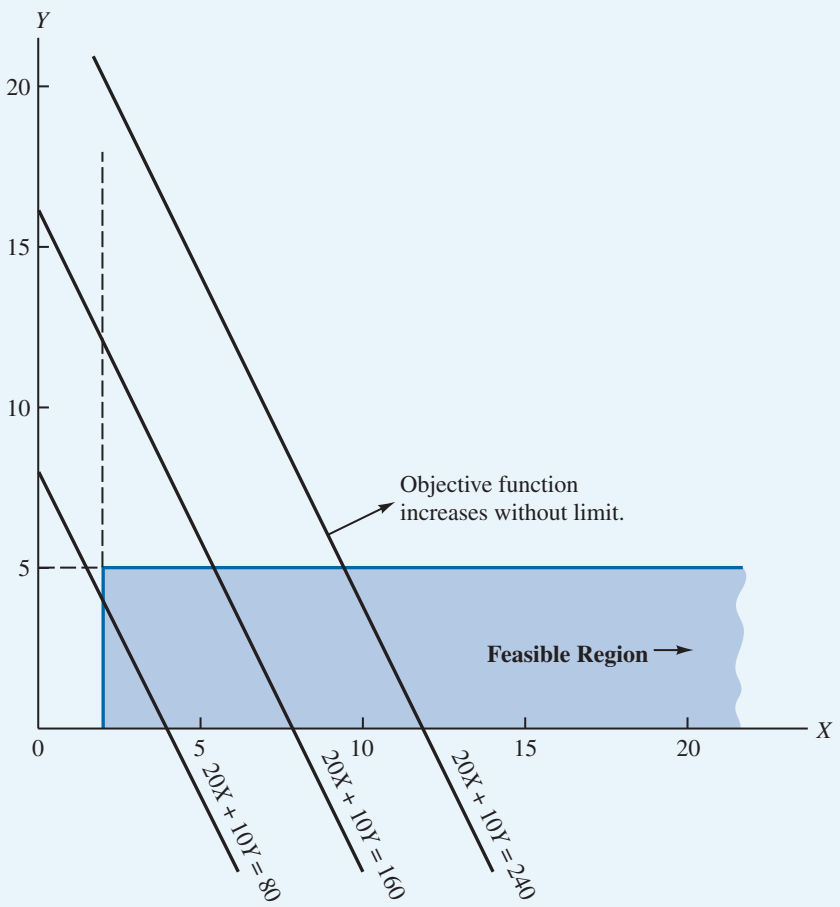
As an illustration, consider the following linear program with two decision variables, X and Y:

$$\begin{aligned} &\text{Max } 20X + 10Y \\ &\text{s.t.} \\ &\quad 1X \qquad\quad \geq 2 \\ &\qquad\quad 1Y \leq 5 \\ &\quad X, Y \qquad\quad \geq 0 \end{aligned}$$

In Figure 14.10 we graph the feasible region associated with this problem. Note that we can indicate only part of the feasible region because the feasible region extends

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.10 Example of an Unbounded Problem



indefinitely in the direction of the *X*-axis. Looking at the objective function lines in Figure 14.10, we see that the solution to this problem may be made as large as we desire. In other words, no matter which solution we pick, we will always be able to reach some feasible solution with a larger value. Thus, we say that the solution to this linear program is *unbounded.*

Whenever you attempt to solve an unbounded problem using Excel Solver, you will receive a message in the Solver Results dialog box telling you that the "Objective Cell values do not converge." In linear programming models of real problems, the occurrence of an unbounded solution means that the problem has been improperly formulated. We know it is not possible to increase profits indefinitely. Therefore, we must conclude that if a profit maximization problem results in an unbounded solution, the mathematical model does not sufficiently represent the real-world problem. In many cases, this error is the result of inadvertently omitting a constraint during problem formulation.

The parameters for optimization models are often less than certain. In the next section, we discuss the sensitivity of the optimal solution to uncertainty in the model parameters. In addition to the optimal solution, Excel Solver can provide some useful information on the sensitivity of that solution to changes in the model parameters.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notes + Comments

1. Infeasibility is independent of the objective function. It exists because the constraints are so restrictive that no feasible region for the linear programming model is possible. Thus, when you encounter infeasibility, making changes in the coefficients of the objective function will not help; the problem will remain infeasible.
2. The occurrence of an unbounded solution is often the result of a missing constraint. However, a change in the objective function may cause a previously unbounded problem to become bounded with an optimal solution. For example, the graph in Figure 14.10 shows an unbounded solution for the objective function Max $20X + 10Y$. However, changing the objective function to Max $-20X - 10Y$ will provide the optimal solution $X = 2$ and $Y = 0$ even though no changes have been made in the constraints.

14.5 Sensitivity Analysis

Sensitivity analysis is the study of how the changes in the input parameters of an optimization model affect the optimal solution. Using sensitivity analysis, we can answer questions such as the following:

1. How will a change *in a coefficient of the objective function* affect the optimal solution?
2. How will a change in the *right-hand-side value for a constraint* affect the optimal solution?

Because sensitivity analysis is concerned with how these changes affect the optimal solution, the analysis does not begin until the optimal solution to the original linear programming problem has been obtained. For that reason, sensitivity analysis is sometimes referred to as *postoptimality analysis.* Let us return to the M&D Chemicals problem as an example of how to interpret the sensitivity report provided by Excel Solver.

Interpreting Excel Solver Sensitivity Report

Recall the M&D Chemicals problem discussed in Section 14.3. We had defined the following decision variables and model:

A = number of gallons of product A
B = number of gallons of product B

$$\begin{aligned} \text{Min} \quad & 2A + 3B \\ \text{s.t.} \quad & \\ & 1A \geq 125 && \text{Demand for product A} \\ & 1A + 1B \geq 350 && \text{Total production} \\ & 2A + 1B \leq 600 && \text{Processing time} \\ & A, B \geq 0 \end{aligned}$$

We found that the optimal solution is $A = 250$ and $B = 100$ with objective function value $= 2(250) + 3(100) = \$800$. The first constraint is not binding, but the second and third constraints are binding because $1(250) + 1(100) = 350$ and $2(250) + 100 = 600$. After running Excel Solver, we may generate the **Sensitivity Report** by selecting **Sensitivity** from the **Reports** section of the **Solver Results** dialog box and then selecting **OK**. The Sensitivity report for the M&D Chemicals problem appears in Figure 14.11. There are two sections in this report: one for decision variables (Variable Cells) and one for Constraints.

Let us begin by interpreting the Constraints section. The cell location of the left-hand side of the constraint, the constraint name, and the value of the left-hand side of the constraint at optimality are given in the first three columns. The fourth column gives the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.11 Solver Sensitivity Report for the M&D Chemicals Problem

Variable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
$B$14	Gallons Produced Product A	250	0	2	1	1E+30
$C$14	Gallons Produced Product B	100	0	3	1E+30	1

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$B$19	Product A Provided	250	0	125	125	1E+30
$B$20	Total Production Provided	350	4	350	125	50
$B$23	Processing Time Hours Used	600	-1	600	100	125

shadow price for each constraint. The **shadow price** for a constraint is the change in the optimal objective function value if the right-hand side of that constraint is increased by one. Let us interpret each shadow price given in the report in Figure 14.11.

The first constraint is: $1A \geq 125$. This is a nonbinding constraint because $250 > 125$. If we change the constraint to $1A \geq 126$, there will be no change in the objective function value. The reason for this is that the constraint will remain nonbinding at the optimal solution, because $1A = 250 > 126$. Hence, the shadow price is zero. In fact, *nonbinding constraints will always have a shadow price of zero.*

The second constraint is binding and its shadow price is 4. The interpretation of the shadow price is as follows. If we change the constraint from $1A + 1B \geq 350$ to $1A + 1B \geq 351$, the optimal objective function value will increase by \$4; that is, the new optimal solution will have an objective function value equal to $\$800 + \$4 = \$804$.

The third constraint is also binding and has a shadow price of -1. The interpretation of the shadow price is as follows. If we change the constraint from $2A + 1B \leq 600$ to $2A + 1B \leq 601$, the objective function value will decrease by \$1; that is, the new optimal solution will have an objective function value equal to $\$800 - \$1 = \$799$.

Note that the shadow price for the second constraint is positive, but for the third it is negative. Why is this? The sign of the shadow price depends on whether the problem is a maximization or a minimization and the type of constraint under consideration. The M&D Chemicals problem is a cost minimization problem. The second constraint is a greater-than-or-equal-to constraint. By increasing the right-hand side, we make the constraint even *more* restrictive. This results in an increase in cost. Contrast this with the third constraint. The third constraint is a less-than-or-equal-to constraint. By increasing the right-hand side, we make more hours available. We have made the constraint less restrictive. Because we have made the constraint *less* restrictive, there are more feasible solutions from which to choose. Therefore, cost drops by \$1.

Making a constraint more restrictive is often referred to as tightening the constraint. Making a constraint less restrictive is often referred to as relaxing, or loosening, the constraint.

When observing shadow prices, the following general principle holds: *Making a binding constraint more restrictive degrades or leaves unchanged the optimal objective function value, and making a binding constraint less restrictive improves or leaves unchanged the optimal objective function.* We shall see several more examples of this later in this chapter. Also, shadow prices are symmetric; so the negative of the shadow price is the change in the objective function for a *decrease* of one in the right-hand side.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In Figure 14.11, the Allowable Increase and the Allowable Decrease are the **Right-hand side allowable increase (decrease)** for which the current shadow price remains valid. For example, because the allowable increase in the processing time is 100, if we increase the processing time hours by 50 to $600 + 50 = 650$, we can say with certainty that the optimal objective function value will change by $(-1)50 = -50$. Hence, we know that the optimal objective function value will be $\$800 - \$50 = \$750$. If we increase the right-hand side of the processing time beyond the allowable increase of 100, we cannot predict what will happen. Likewise, if we decrease the right-hand side of the processing time constraint by 50, we know that the optimal objective function value will change by the negative of the shadow price: $-(-1)50 = 50$. Cost will increase by $50. If we change the right-hand side by more than the allowable increase or decrease, the shadow price is no longer valid.

Let us now turn to the Variable Cells section of the Sensitivity Report. As in the constraint section, the cell location, variable name, and final (optimal) value for each variable are given. The fourth column is Reduced Cost. The **reduced cost** for a decision variable is the shadow price of the nonnegativity constraint for that variable. In other words, the reduced cost indicates the change in the optimal objective function value that results from changing the right-hand side of the nonnegativity constraint from 0 to 1.

In the fifth column of the report, the objective function coefficient for the variable is given. The Allowable Increase and Allowable Decrease indicate the **objective function coefficient allowable increase and decrease** for which the *current optimal solution will remain optimal*. The value 1E + 30 in the report is essentially infinity. So long as the cost of product A is greater than or equal to negative infinity and less than or equal to $2 + 1 = 3$, the current solution remains optimal. For example, if the cost of product A is really $2.50 per gallon, we do not need to re-solve the model. Because the increase in cost of $0.50 is less than the allowable increase of $1.00, the current solution of 250 gallons of product A and 100 gallons of product B remains optimal.

As we have seen, the Excel Solver Sensitivity Report can provide useful information about the sensitivity of the optimal solution to changes in the model input data. However, this type of classical sensitivity analysis is somewhat limited. Classical sensitivity analysis is based on the assumption that only one piece of input data has changed; it is assumed that all other parameters remain as stated in the original problem. In many cases, however, we are interested in what would happen if two or more pieces of input data are changed simultaneously. The easiest way to examine the effect of simultaneous changes is to make the changes and re-solve the model.

Notes + Comments

We defined the reduced cost as the shadow price of the nonnegativity constraint for that variable. When there is a binding simple upper-bound constraint for a variable, the reduced cost reported by Excel Solver is the shadow price of that upper-bound constraint. Likewise, if there is a binding nonzero lower bound for a variable, the reduced cost is the shadow price for that lower-bound constraint. So to be more general, the reduced cost for a decision variable is the shadow price of the binding simple lower- or upper-bound constraint for that variable.

14.6 General Linear Programming Notation and More Examples

Earlier in this chapter, we showed how to formulate linear programming models for the Par, Inc. and M&D Chemicals problems. To formulate a linear programming model of the Par, Inc. problem, we began by defining two decision variables: S = number of standard bags and D = number of deluxe bags. In the M&D Chemicals problem, the two decision variables were defined as A = number of gallons of product A and B = number of gallons of product B. We selected decision-variable names of S and D

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

in the Par, Inc. problem and A and B in the M&D Chemicals problem to make it easier to recall what these decision variables represented in the problem. Although this approach works well for linear programs involving a small number of decision variables, it can become difficult when dealing with problems involving a large number of decision variables.

A more general notation that is often used for linear programs uses the letter x with a subscript. For instance, in the Par, Inc. problem, we could have defined the decision variables as follows:

$$x_1 = \text{number of standard bags}$$
$$x_2 = \text{number of deluxe bags}$$

In the M&D Chemicals problem, the same variable names would be used, but their definitions would change:

$$x_1 = \text{number of gallons of product A}$$
$$x_2 = \text{number of gallons of product B}$$

A disadvantage of using general notation for decision variables is that we are no longer able to easily identify what the decision variables actually represent in the mathematical model. However, the advantage of general notation is that formulating a mathematical model for a problem that involves a large number of decision variables is much easier. For instance, for a linear programming model with three decision variables, we would use variable names of x_1, x_2, and x_3; for a problem with four decision variables, we would use variable names of x_1, x_2, x_3, and x_4; and so on. Clearly, if a problem involved 1,000 decision variables, trying to identify 1,000 unique names would be difficult. However, using the general linear programming notation, the decision variables would be defined as $x_1, x_2, x_3, \ldots, x_{1000}$.

Using this new general notation, the Par, Inc. model would be written as follows:

$$\begin{aligned}
&\text{Max } 10x_1 + 9x_2 \\
&\text{s.t.} \\
&\tfrac{7}{10}x_1 + 1x_2 \le 630 && \text{Cutting and dyeing} \\
&\tfrac{1}{2}x_1 + \tfrac{5}{6}x_2 \le 600 && \text{Sewing} \\
&1x_1 + \tfrac{2}{3}x_2 \le 708 && \text{Finishing} \\
&\tfrac{1}{10}x_1 + \tfrac{1}{4}x_2 \le 135 && \text{Inspection and packaging} \\
&x_1, x_2 \ge 0
\end{aligned}$$

In some of the examples that follow in this section and in Chapters 15 and 16, we will use this type of subscripted notation.

Investment Portfolio Selection

In finance, linear programming can be applied in problem situations involving capital budgeting, make-or-buy decisions, asset allocation, portfolio selection, financial planning, and many more. Next, we describe a portfolio selection problem.

Portfolio selection problems involve situations in which a financial manager must select specific investments—for example, stocks and bonds—from a variety of investment alternatives. Managers of mutual funds, credit unions, insurance companies, and banks frequently encounter this type of problem. The objective function for portfolio selection problems usually is maximization of expected return or minimization of risk. The constraints usually take the form of restrictions on the type of permissible investments, state laws, company policy, maximum permissible risk, and so on. Problems of this type have been formulated and solved using a variety of optimization techniques. In this section, we formulate and solve a portfolio selection problem as a linear program.

Consider the case of Welte Mutual Funds, Inc. located in New York City. Welte just obtained $100,000 by converting industrial bonds to cash and is now looking for other

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

investment opportunities for these funds. Based on Welte's current investments, the firm's top financial analyst recommends that all new investments be made in the oil industry, steel industry, or government bonds. Specifically, the analyst identified five investment opportunities and projected their annual rates of return. The investments and rates of return are shown in Table 14.3.

The management at Welte imposed the following investment guidelines:

1. Neither industry (oil or steel) should receive more than $50,000.
2. The amount invested in government bonds should be at least 25% of the steel industry investments.
3. The investment in Pacific Oil, the high-return but high-risk investment, cannot be more than 60% of the total oil industry investment.

What portfolio recommendations—investments and amounts—should be made for the available $100,000? Given the objective of maximizing projected return subject to the budgetary and managerially imposed constraints, we can answer this question by formulating and solving a linear programming model of the problem. The solution will provide investment recommendations for the management of Welte Mutual Funds.

Let us define the following decision variables:

x_1 = dollars invested in Atlantic Oil
x_2 = dollars invested in Pacific Oil
x_3 = dollars invested in Midwest Steel
x_4 = dollars invested in Huber Steel
x_5 = dollars invested in government bonds

Using the projected rates of return shown in Table 14.3, we write the objective function for maximizing the total return for the portfolio as

$$\text{Max} \quad 0.073x_1 + 0.103x_2 + 0.064x_3 + 0.075x_4 + 0.045x_5$$

The constraint specifying investment of the available $100,000 is

$$x_1 + x_2 + x_3 + x_4 + x_5 = 100{,}000$$

The requirements that neither the oil nor steel industry should receive more than $50,000 are as follows

$$x_1 + x_2 \leq 50{,}000$$
$$x_3 + x_4 \leq 50{,}000$$

The requirement that the amount invested in government bonds be at least 25% of the steel industry investment is expressed as

$$x_5 \geq 0.25(x_3 + x_4)$$

Finally, the constraint that Pacific Oil cannot be more than 60% of the total oil industry investment is

$$x_2 \leq 0.60(x_1 + x_2)$$

Table 14.3 Investment Opportunities for Welte Mutual Funds

Investment	Projected Rate of Return (%)
Atlantic Oil	7.3
Pacific Oil	10.3
Midwest Steel	6.4
Huber Steel	7.5
Government bonds	4.5

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

By adding the nonnegativity restrictions, we obtain the complete linear programming model for the Welte Mutual Funds investment problem:

$$\text{Max} \quad 0.073x_1 + 0.103x_2 + 0.064x_3 + 0.075x_4 + 0.045x_5$$

s.t.

$$\begin{aligned}
&x_1 + x_2 + x_3 + x_4 + x_5 = 100{,}000 && \text{Available funds}\\
&x_1 + x_2 \le 50{,}000 && \text{Oil industry maximum}\\
&x_3 + x_4 \le 50{,}000 && \text{Steel industry maximum}\\
&x_5 \ge 0.25(x_3 + x_4) && \text{Government bonds minimum}\\
&x_2 \le 0.60(x_1 + x_2) && \text{Pacific Oil restriction}\\
&x_1, x_2, x_3, x_4, x_5 \ge 0
\end{aligned}$$

The optimal solution to this linear program is shown in Figure 14.12. Note that the optimal solution indicates that the portfolio should be diversified among all the investment opportunities except Midwest Steel. The projected annual return for this portfolio is $8,000, which is an overall return of 8%. Except for the upper bound on the Steel investment, all constraints are binding.

Figure 14.12 The Solution for the Welte Mutual Funds Problem

	A	B	C	D	E	F
1	**Welte Mutual Funds Problem**					
2						
3	**Parameters**					
4	Investment	Projected Rate of Return				
5	Atlantic Oil	0.073		Available Funds	100000	
6	Pacific Oil	0.103		Oil Max	50000	
7	Midwest Steel	0.064		Steel Max	50000	
8	Huber Steel	0.075		Pacific Oil Max	0.6	
9	Gov't Bonds	0.045		Gov't Bonds Min	0.25	
10						
11	**Model**					
12						
13	Investment	Amount Invested				
14	Atlantic Oil	20000				
15	Pacific Oil	30000				
16	Midwest Steel	0				
17	Huber Steel	40000				
18	Gov't Bonds	10000				
19						
20	Max Total Return	=SUMPRODUCT(B5:B9,B14:B18)				
21						
22		Funds Invested	Funds Available	Unused Funds		
23	Total	=SUM(B14:B18)	=E5	=C23-B23		
24						
25		Funds Invested	Max Allowed			
26	Oil	=SUM(B14:B15)	=E6			
27	Steel	=SUM(B16:B17)	=E7			
28	Pacific Oil	=B15	=E8*(B14+B15)			
29						
30		Funds Invested	Min Required			
31	Gov't Bonds	=B18	=E9*(B16+B17)			

MODEL file
welte

	A	B	C	D	E
1	**Welte Mutual Funds Problem**				
2					
3	**Parameters**				
4	Investment	Projected Rate of Return			
5	Atlantic Oil	0.073		Available Funds	$100,000.00
6	Pacific Oil	0.103		Oil Max	$50,000.00
7	Midwest Steel	0.064		Steel Max	$50,000.00
8	Huber Steel	0.075		Pacific Oil Max	0.6
9	Gov't Bonds	0.045		Gov't Bonds Min	0.25
10					
11	**Model**				
12					
13	Investment	Amount Invested			
14	Atlantic Oil	$20,000.00			
15	Pacific Oil	$30,000.00			
16	Midwest Steel	$0.00			
17	Huber Steel	$40,000.00			
18	Gov't Bonds	$10,000.00			
19					
20	Max Total Return	$8,000.00			
21					
22		Funds Invested	Funds Available	Unused Funds	
23	Total	$100,000.00	$100,000.00	$0.00	
24					
25		Funds Invested	Max Allowed		
26	Oil	$50,000.00	$50,000.00		
27	Steel	$40,000.00	$50,000.00		
28	Pacific Oil	$30,000.00	$30,000.00		
29					
30		Funds Invested	Min Required		
31	Gov't Bonds	$10,000.00	$10,000.00		

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notes + Comments

1. The optimal solution to the Welte Mutual Funds problem indicates that $20,000 should be spent on the Atlantic Oil stock. If Atlantic Oil sells for $75 per share, we would have to purchase exactly 266⅔ shares in order to spend exactly $20,000. The difficulty of purchasing fractional shares can be handled by purchasing the largest possible integer number of shares with the allotted funds (e.g., 266 shares of Atlantic Oil). This approach guarantees that the budget constraint will not be violated. This approach, of course, introduces the possibility that the solution will no longer be optimal, but the danger is slight if a large number of securities are involved. In cases in which the analyst believes that the decision variables *must* have integer values, the problem must be formulated as an integer linear programming model (the topic of Chapter 15).
2. Financial portfolio theory stresses obtaining a proper balance between risk and return. In the Welte problem, we explicitly considered return in the objective function. Risk is controlled by choosing constraints that ensure diversity among oil and steel stocks and a balance between government bonds and the steel industry investment. In Chapter 16, we discuss investment portfolio models that control risk as measured by the variance of returns on investment.

Transportation Planning

The **transportation problem** arises frequently in planning for the distribution of goods and services from several supply locations to several demand locations. Typically, the quantity of goods available at each supply location (origin) is limited, and the quantity of goods needed at each of several demand locations (destinations) is known. The usual objective in a transportation problem is to minimize the cost of shipping goods from the origins to the destinations.

Let us revisit the transportation problem faced by Foster Generators, discussed in Chapter 12. This problem involves the transportation of a product from three plants to four distribution centers. Foster Generators operates plants in Cleveland, Ohio; Bedford, Indiana; and York, Pennsylvania. Production capacities over the next three-month planning period for one type of generator are as follows:

Origin	Plant	Three-Month Production Capacity (units)
1	Cleveland	5,000
2	Bedford	6,000
3	York	2,500
	Total	13,500

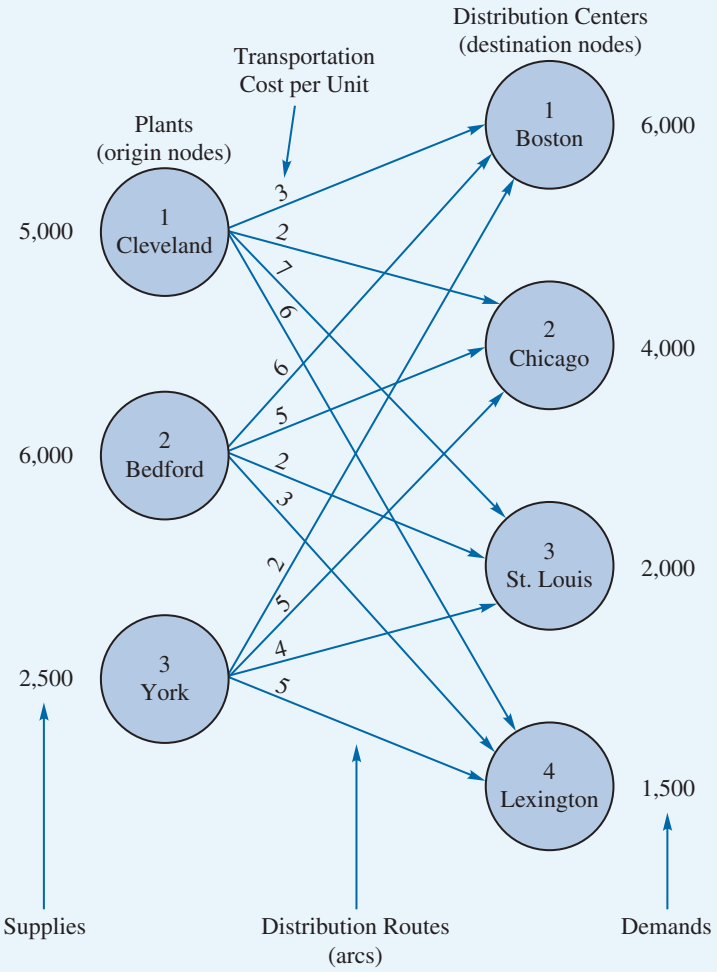
The firm distributes its generators through four regional distribution centers located in Boston, Massachusetts; Chicago, Illinois; St. Louis, Missouri; and Lexington, Kentucky; the three-month forecast of demand for the distribution centers is as follows:

Destination	Distribution Center	Three-Month Demand Forecast (units)
1	Boston	6,000
2	Chicago	4,000
3	St. Louis	2,000
4	Lexington	1,500
	Total	13,500

Management would like to determine how much of its production should be shipped from each plant to each distribution center. Figure 14.13 shows graphically the 12 distribution routes Foster can use. Such a graph is called a *network*; the circles are referred to as *nodes*, and the lines connecting the nodes as *arcs*. Each origin and each destination is represented by a node, and each possible shipping route is represented by an arc. The amount of the supply

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.13 The Network Representation of the Foster Generators Transportation Problem



is written next to each origin node, and the amount of the demand is written next to each destination node. The goods shipped from the origins to the destinations represent the flow in the network. Note that the direction of flow (from origin to destination) is indicated by the arrows.

For Foster's transportation problem, the objective is to determine the routes to be used and the quantity to be shipped via each route that will provide the minimum total transportation cost. The cost for each unit shipped on each route is given in Table 14.4 and is shown on each arc in Figure 14.13.

A linear programming model can be used to solve this transportation problem. We use *double-subscripted* decision variables, with x_{11} denoting the number of units shipped from origin 1 (Cleveland) to destination 1 (Boston), x_{12} denoting the number of units shipped from origin 1 (Cleveland) to destination 2 (Chicago), and so on. In general, the decision variables for a transportation problem having m origins and n destinations are written as follows:

$$x_{ij} = \text{number of units shipped from origin } i \text{ to destination } j$$
$$\text{where } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 14.4 Transportation Cost per Unit for the Foster Generators Transportation Problem ($)

	Destination			
Origin	Boston	Chicago	St. Louis	Lexington
Cleveland	3	2	7	6
Bedford	6	5	2	3
York	2	5	4	5

Because the objective of the transportation problem is to minimize the total transportation cost, we can use the cost data in Table 14.4 or on the arcs in Figure 14.13 to develop the following cost expressions:

$$\begin{aligned}
&\text{Transportation costs for units shipped from Cleveland} = 3x_{11} + 2x_{12} + 7x_{13} + 6x_{14}\\
&\text{Transportation costs for units shipped from Bedford} = 6x_{21} + 5x_{22} + 2x_{23} + 3x_{24}\\
&\text{Transportation costs for units shipped from York} = 2x_{31} + 5x_{32} + 4x_{33} + 5x_{34}
\end{aligned}$$

The sum of these expressions provides the objective function showing the total transportation cost for Foster Generators.

Transportation problems need constraints because each origin has a limited supply and each destination has a demand requirement. We consider the supply constraints first. The capacity at the Cleveland plant is 5,000 units. With the total number of units shipped from the Cleveland plant expressed as $x_{11} + x_{12} + x_{13} + x_{14}$, the supply constraint for the Cleveland plant is

$$x_{11} + x_{12} + x_{13} + x_{14} \leq 5{,}000 \qquad \text{Cleveland supply}$$

With three origins (plants), the Foster transportation problem has three supply constraints. Given the capacity of 6,000 units at the Bedford plant and 2,500 units at the York plant, the two additional supply constraints are as follows:

$$\begin{aligned}
x_{21} + x_{22} + x_{23} + x_{24} &\leq 6{,}000 \quad \text{Bedford supply}\\
x_{31} + x_{32} + x_{33} + x_{34} &\leq 2{,}500 \quad \text{York supply}
\end{aligned}$$

With the four distribution centers as the destinations, four demand constraints are needed to ensure that destination demands will be satisfied:

$$\begin{aligned}
x_{11} + x_{21} + x_{31} &= 6{,}000 \quad \text{Boston demand}\\
x_{12} + x_{22} + x_{32} &= 4{,}000 \quad \text{Chicago demand}\\
x_{13} + x_{23} + x_{33} &= 2{,}000 \quad \text{St. Louis demand}\\
x_{14} + x_{24} + x_{34} &= 1{,}500 \quad \text{Lexington demand}
\end{aligned}$$

Combining the objective function and constraints into one model provides a 12-variable, 7-constraint linear programming formulation of the Foster Generators transportation problem:

$$\begin{aligned}
\text{Min} \quad & 3x_{11} + 2x_{12} + 7x_{13} + 6x_{14} + 6x_{21} + 5x_{22} + 2x_{23} + 3x_{24} + 2x_{31} + 5x_{32} + 4x_{33} + 5x_{34}\\
\text{s.t.}\\
& x_{11} + x_{12} + x_{13} + x_{14} \leq 5{,}000\\
& x_{21} + x_{22} + x_{23} + x_{24} \leq 6{,}000\\
& x_{31} + x_{32} + x_{33} + x_{34} \leq 2{,}500\\
& x_{11} + x_{21} + x_{31} = 6{,}000\\
& x_{12} + x_{22} + x_{32} = 4{,}000\\
& x_{13} + x_{23} + x_{33} = 2{,}000\\
& x_{14} + x_{24} + x_{34} = 1{,}500\\
& x_{ij} \geq 0 \qquad \text{for } i = 1, 2, 3 \text{ and } j = 1, 2, 3, 4
\end{aligned}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Comparing the linear programming formulation to the network in Figure 14.13 leads to several observations. All the information needed for the linear programming formulation is on the network. Each node has one constraint, and each arc has one variable. The sum of the variables corresponding to arcs from an origin node must be less than or equal to the origin's supply, and the sum of the variables corresponding to the arcs into a destination node must be equal to the destination's demand.

A spreadsheet model and the solution to the Foster Generators problem (Figure 14.14) show that the minimum total transportation cost is $39,500. The values for the decision variables show the optimal amounts to ship over each route. For example, 1,000 units should be shipped from Cleveland to Boston and 4,000 units should be shipped from Cleveland to Chicago. Other values of the decision variables indicate the remaining shipping quantities and routes.

Figure 14.14 Spreadsheet Model and Solution for the Foster Generator Problem

	A	B	C	D	E	F
1	**Foster Generators**					
2	**Parameters**					
3	Shipping Cost/Unit		Destination			
4	Origin	Boston	Chicago	St. Louis	Lexington	Supply
5	Cleveland	3	2	7	6	5000
6	Bedford	6	5	2	3	6000
7	York	2	5	4	5	2500
8	Demand	6000	4000	2000	1500	
9						
10						
11	**Model**					
12						
13	Total Cost	=SUMPRODUCT(B5:E7,B17:E19)				
14						
15			Destination			
16	Origin	Boston	Chicago	St. Louis	Lexington	**Total**
17	Cleveland	1000	4000	0	0	=SUM(B17:E17)
18	Bedford	2500	0	2000	1500	=SUM(B18:E18)
19	York	2500	0	0	0	=SUM(B19:E19)
20	Total	=SUM(B17:B19)	=SUM(C17:C19)	=SUM(D17:D19)	=SUM(E17:E19)	

MODEL *file*
foster

	A	B	C	D	E	F
1	**Foster Generators**					
2	**Parameters**					
3	Shipping Cost/Unit		Destination			
4	Origin	Boston	Chicago	St. Louis	Lexington	Supply
5	Cleveland	$3.00	$2.00	$7.00	$6.00	5000
6	Bedford	$6.00	$5.00	$2.00	$3.00	6000
7	York	$2.00	$5.00	$4.00	$5.00	2500
8	Demand	6000	4000	2000	1500	
9						
10						
11	**Model**					
12						
13	Total Cost	$39,500.00				
14						
15			Destination			
16	Origin	Boston	Chicago	St. Louis	Lexington	**Total**
17	Cleveland	1000	4000	0	0	5000
18	Bedford	2500	0	2000	1500	6000
19	York	2500	0	0	0	2500
20	Total	6000	4000	2000	1500	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The network structure shown in Figure 14.13 gives the solution to a transportation problem a very special property. So long as the supplies and demands are integers, the optimal values of the decision variable will also be integer. That is, as long as the supplies and demands are integers, the optimal shipments will never be fractional.

Maximizing Banner Ad Revenue

Applications of linear programming to marketing are numerous. Advertising campaign mix, marketing mix, and marketing research are just a few areas of application. In this section, we consider the allocation of web site banner ads.

Content publishers such as CNN, Fox News, MSNBC, and Yahoo generate revenue by placing banner ads on their web sites. These publishers are paid based on the number of "click-throughs" on banner ads, that is, the number of times someone clicks the banner ad and goes on to the linked web site of the company that is advertising. For a given day, publishers need to determine how to allocate banner ad showings, also known as impressions, across their content pages for each advertiser so as to maximize banner ad revenue.

Let us consider the Business, Science, and Sports sections of a large content publisher, MHT. MHT has contracts for banner ads with five companies, ecommerce company Nile, management consulting company Zstar, athletic wear company Cheetah, telecommunications firm Stride, and home goods retailer Stove. Based on historical data from past contracts, MHT has determined the expected click-through rates for each of these advertisers by section of their web content. The expected click-through rates are given in the upper portion of Table 14.5. For example, for every 10,000 impressions of a banner ad for Nile in the Business section, $10{,}000 \times 0.0155 = 155$ click-throughs to the Nile web site are expected to occur. The other click-through rates given in Table 14.5 are interpreted in a similar way.

MHT contracts with each of the five companies include a minimum and maximum number of impressions. The minimum and maximum number of impressions are provided in the lower portion of Table 14.5.

For next Friday, MHT expects to have sufficient visitors to its web site to guarantee 2,550,000 visitors to its Business section, 2,150,000 visitors to its Science section, and 2,500,000 visitors to its Sports section. MHT receives $0.30 for each click-through achieved, regardless of company or section from which the click-through originates.

MHT advertising planners need to determine, for next Friday, how to allocate impressions for each of the five advertisers across the three sections of the MHT web site. A linear programming model, similar to the Foster Generators problem model

Table 14.5 Data for the MHT Banner Ad Allocation Problem

Click-Through Rates

Section	Nile	Zstar	Cheetah	Stride	Stove
Business	0.0155	0.0265	0.0100	0.0170	0.0105
Science	0.0165	0.0110	0.0125	0.0265	0.0125
Sports	0.0145	0.0235	0.0190	0.0225	0.0160

Contract Impression Limits

	Nile	Zstar	Cheetah	Stride	Stove
Ads Lower Limit	1,500,000	1,000,000	1,100,000	1,000,000	1,500,000
Ads Upper Limit	2,000,000	2,000,000	1,800,000	2,000,000	2,000,000

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

previously discussed, can be used to allocate impressions so as to maximize the revenue received by MHT.

Let us define decision variables as follows:

$$x_{ij} = \text{the number of impressions in section } i \text{ allocated to company } j$$

where $i = 1$ is Business, $i = 2$ is Science, $i = 3$ is Sports and $j = 1$ is Nile, $j = 2$ is Zstar, $j = 3$ is Cheetah, $j = 4$ is Stride, and $j = 5$ is Stove.

The objective function is to maximize expected revenue, which can be calculated by multiplying the per-click-through revenue, the expected click-through rate and the allocated impressions. For example, for Business section impressions for Cheetah, we have $(0.30)(0.01)\, x_{13} = 0.003x_{13}$. The expected revenues from the three sections are:

$$\begin{aligned}
&0.3(0.0155x_{11} + 0.0265x_{12} + 0.010x_{13} + 0.0170x_{14} + 0.0105x_{15}) && \text{(Business)}\\
&0.3(0.0165x_{21} + 0.0110x_{22} + 0.0125x_{23} + 0.0265x_{24} + .0125x_{25}) && \text{(Science)}\\
&0.3(0.0145x_{31} + 0.0235x_{32} + 0.0190x_{33} + 0.0225x_{34} + 0.0160x_{35}) && \text{(Sports)}
\end{aligned}$$

The objective function is the sum of the expected revenues generated from the three sections. Multiplying and summing across the three sections we have:

$$\begin{aligned}
\text{Maximize} \quad & 0.0047x_{11} + 0.0080x_{12} + 0.0030x_{13} + 0.0051x_{14} + 0.0032x_{15} + 0.0050x_{21}\\
& + 0.0033x_{22} + 0.0038x_{23} + 0.0008x_{24} + 0.0038x_{25} + 0.0044x_{31}\\
& + 0.0071x_{32} + 0.0057x_{33} + .0068x_{34} + 0.0048x_{35}
\end{aligned}$$

Now let us consider the constraints of the problem. Since each section is limited to the number of impressions that MHT can guarantee, we have the following impression availability constraints:

$$\begin{aligned}
&x_{11} + x_{12} + x_{13} + x_{14} + x_{15} \le 2{,}550{,}000 && \text{(Business Section Impressions Available)}\\
&x_{21} + x_{22} + x_{23} + x_{24} + x_{25} \le 2{,}150{,}000 && \text{(Science Section Impressions Available)}\\
&x_{31} + x_{32} + x_{33} + x_{34} + x_{35} \le 2{,}500{,}000 && \text{(Sports Section Impressions Available)}
\end{aligned}$$

Each company has a lower limit and an upper limit on impressions received:

$$\begin{aligned}
&x_{11} + x_{21} + x_{31} \ge 1{,}500{,}000 && \text{(Lower Limit for Nile)}\\
&x_{12} + x_{22} + x_{32} \ge 1{,}000{,}000 && \text{(Lower Limit for Zstar)}\\
&x_{13} + x_{23} + x_{33} \ge 1{,}100{,}000 && \text{(Lower Limit for Cheetah)}\\
&x_{14} + x_{24} + x_{34} \ge 1{,}000{,}000 && \text{(Lower Limit for Stride)}\\
&x_{15} + x_{25} + x_{35} \ge 1{,}500{,}000 && \text{(Lower Limit for Stove)}
\end{aligned}$$

$$\begin{aligned}
&x_{11} + x_{21} + x_{31} \le 2{,}000{,}000 && \text{(Upper Limit for Nile)}\\
&x_{12} + x_{22} + x_{32} \le 2{,}000{,}000 && \text{(Upper Limit for Zstar)}\\
&x_{13} + x_{23} + x_{33} \le 1{,}800{,}000 && \text{(Upper Limit for Cheetah)}\\
&x_{14} + x_{24} + x_{34} \le 2{,}000{,}000 && \text{(Upper Limit for Stride)}\\
&x_{15} + x_{25} + x_{35} \le 2{,}000{,}000 && \text{(Upper Limit for Stove)}
\end{aligned}$$

Adding constraints that ensure variables must be nonnegative, we have the following linear program:

$$\begin{aligned}
\text{Maximize} \quad & 0.0047x_{11} + 0.0080x_{12} + 0.0030x_{13} + 0.0051x_{14} + 0.0032x_{15} + 0.0050x_{21}\\
& + 0.0033x_{22} + 0.0038x_{23} + 0.0008x_{24} + 0.0038x_{25} + 0.0044x_{31}\\
& + 0.0071x_{32} + 0.0057x_{33} + 0.0068x_{34} + 0.0048x_{35}
\end{aligned}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

s.t.

$$
\begin{aligned}
x_{11} + x_{12} + x_{13} + x_{14} + x_{15} &\le 2{,}550{,}000 && \text{(Business Section Impressions Available)}\\
x_{21} + x_{22} + x_{23} + x_{24} + x_{25} &\le 2{,}150{,}000 && \text{(Science Section Impressions Available)}\\
x_{31} + x_{32} + x_{33} + x_{34} + x_{35} &\le 2{,}500{,}000 && \text{(Sports Impressions Available)}\\
x_{11} + x_{21} + x_{31} &\ge 1{,}500{,}000 && \text{(Lower Limit for Nile)}\\
x_{12} + x_{22} + x_{32} &\ge 1{,}000{,}000 && \text{(Lower Limit for Zstar)}\\
x_{13} + x_{23} + x_{33} &\ge 1{,}100{,}000 && \text{(Lower Limit for Cheetah)}\\
x_{14} + x_{24} + x_{34} &\ge 1{,}000{,}000 && \text{(Lower Limit for Stride)}\\
x_{15} + x_{25} + x_{35} &\ge 1{,}500{,}000 && \text{(Lower Limit for Stove)}\\
x_{11} + x_{21} + x_{31} &\le 2{,}000{,}000 && \text{(Upper Limit for Nile)}\\
x_{12} + x_{22} + x_{32} &\le 2{,}000{,}000 && \text{(Upper Limit for Zstar)}\\
x_{13} + x_{23} + x_{33} &\le 1{,}800{,}000 && \text{(Upper Limit for Cheetah)}\\
x_{14} + x_{24} + x_{34} &\le 2{,}000{,}000 && \text{(Upper Limit for Stride)}\\
x_{15} + x_{25} + x_{35} &\le 2{,}000{,}000 && \text{(Upper Limit for Stove)}\\
x_{ij} \ge 0 \text{ for } i = 1, 2, 3 \text{ and } j &= 1, 2, 3, 4, 5 && \text{(Nonnegativity)}
\end{aligned}
$$

A spreadsheet model and an optimal solution to this linear programming problem are shown in Figure 14.15. The solution indicates that all available impressions should be used, and the resulting revenue is $45,270. Nile impressions are split between the Business and Science sections and Stove impressions are split over the Science and Sports sections. Zstar, Cheetah and Stride are allocated exclusively to the Business, Sports and Science sections, respectively.

The sensitivity report for the MHT solution is shown in Figure 14.16. The nonzero reduced costs are all negative, indicating how much expected revenue will drop per impression that deviates from the plan. So, for example, if we use an impression for Cheetah in the Business section (row 11), then expected revenue will drop by $0.0014. Viewed another way, if the expected revenue per impression increases by at least $0.0014 then this alternative becomes attractive and would be in the solution. Note that the coefficient in the objective function for impressions in the Business section for Cheetah (the objective function coefficient of x_{13}) is ($0.30)(0.01) = $0.003. Assuming the revenue per click-through of $0.30 is fixed, then we can interpret the reduced cost as how much the click-through rate would have to improve before that alternative becomes attractive enough to be in the solution. Hence, if a better placement, or a more attractive banner ad, could increase the click-through rate to at least 0.01 + 0.0014 = 0.0114, then it would be optimal to use that alternative. Other nonzero reduced costs are interpreted in the same way. The allowable increase and decrease in rows 9 through 23 indicate by how much the expected revenue per impression can change and the current solution remains optimal. So, for example, based on row 9, so long as the expected revenue per impression is between 0.00465 − 0.0003 = 0.00435 and 0.00465 + 0.0003 = 0.00495, then the current solution remains optimal.

Next, consider the Constraints section of Figure 14.16. Rows 28 through 32 correspond to the upper limit for each company's impressions. Only Zstar's upper limit is binding (the upper limit is 2,000,000 and 2E + 06 = 2,000,000 impressions are allocated). The shadow price of $0.0003 in row 29 indicates that for every increase of one in the upper limit on impressions, revenue will increase by $0.0003. This will be true for an increase of up to 100,000 (Allowable Increase of 100,000).

Rows 33 through 37 in Figure 14.16 correspond to the lower limits on each company's impressions. Nile, Cheetah and Stove are each at their lower limit and have nonzero shadow prices. As expected, each of these shadow prices is negative, indicating that if we raise the lower limit, it will have a negative impact on revenue. For example, raising the lower limit on Nile (row 33) from 1,500,000 to 1,500,001 will decrease revenue by $0.003.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.15 A Spreadsheet Model and the Solution to the MHT Banner Ad Problem

	A	B	C	D	E	F	G
1	**MHT**						
2	**Parameters**						
3	Section	Nile	Zstar	Cheetah	Stride	Stove	Impressions Available
4	Business	0.0155	0.0265	0.01	0.017	0.0105	2550000
5	Science	0.0165	0.011	0.0125	0.0265	0.0125	2150000
6	Sports	0.0145	0.0235	0.019	0.0225	0.016	2500000
7	Contracted Ads Lower Limit	1500000	1000000	1100000	1000000	1500000	
8	Contracted Ads Upper Limit	2000000	2000000	1800000	2000000	2000000	
9							
10	Revenue per click-through	0.3					
11							
12	Revenue per view	Nile	Zstar	Cheetah	Stride	Stove	
13	Business	=B4*$B$10	=C4*$B$10	=D4*$B$10	=E4*$B$10	=F4*$B$10	
14	Science	=B5*$B$10	=C5*$B$10	=D5*$B$10	=E5*$B$10	=F5*$B$10	
15	Sports	=B6*$B$10	=C6*$B$10	=D6*$B$10	=E6*$B$10	=F6*$B$10	
16							
17	**Model**						
18			Number of Banner Ads				
19		Nile	Zstar	Cheetah	Stride	Stove	
20	Business	550000	2000000	0	0	0	=SUM(B20:F20)
21	Science	950000	0	0	1100000	100000	=SUM(B21:F21)
22	Sports	0	0	1100000	0	1400000	=SUM(B22:F22)
23		=SUM(B20:B22)	=SUM(C20:C22)	=SUM(D20:D22)	=SUM(E20:E22)	=SUM(F20:F22)	
24							
25	Click-Through Revenue	=SUMPRODUCT(B13:F15,B20:F22)					

MODEL *file*
mht

	A	B	C	D	E	F	G
1	**MHT**						
2	**Parameters**						
3	Section	Nile	Zstar	Cheetah	Stride	Stove	Impressions Available
4	Business	0.0155	0.0265	0.0100	0.0170	0.0105	2,550,000
5	Science	0.0165	0.0110	0.0125	0.0265	0.0125	2,150,000
6	Sports	0.0145	0.0235	0.0190	0.0225	0.0160	2,500,000
7	Contracted Ads Lower Limit	1,500,000	1,000,000	1,100,000	1,000,000	1,500,000	
8	Contracted Ads Upper Limit	2,000,000	2,000,000	1,800,000	2,000,000	2,000,000	
9							
10	Revenue per click-through	$0.30					
11							
12	Revenue per view	Nile	Zstar	Cheetah	Stride	Stove	
13	Business	$0.0047	$0.0080	$0.0030	$0.0051	$0.0032	
14	Science	$0.0050	$0.0033	$0.0038	$0.0080	$0.0038	
15	Sports	$0.0044	$0.0071	$0.0057	$0.0068	$0.0048	
16							
17	**Model**						
18				Number of Banner Ads			
19		Nile	Zstar	Cheetah	Stride	Stove	
20	Business	550000	2000000	0	0	0	2550000
21	Science	950000	0	0	1100000	100000	2150000
22	Sports	0	0	1100000	0	1400000	2500000
23		1500000	2000000	1100000	1100000	1500000	
24							
25	Click-Through Revenue	$45,270.00					

This is true for an increase of up to 100,000 (Allowable Increase). Other nonzero shadow prices in rows 33 through 37 are interpreted in the same way.

Rows 38 through 40 pertain to the limits of available impressions for each of the three sections of the MHT web site. Note that all three shadow prices are positive. This is because each of these constraints is binding, and an increase in the right-hand side allows more overall impressions. More allowed impressions will increase revenue. Of the three web site sections, the Sports section (row 40) has the largest positive shadow price ($0.009), meaning that every additional impression beyond 2,500,000 up to an increase of 100,000 (Allowable Increase) will increase revenue by $0.009; the shadow prices for Business and Science in rows 38 and 39 are interpreted in the same way.

As we have seen, the Excel Solver Sensitivity Report can provide useful information about the sensitivity of the optimal solution to changes in the model input. However, this type of classical sensitivity analysis is somewhat limited. Classical sensitivity analysis assumes that only one piece of input data has changed; it assumes that all other parameters remain as stated in the original problem. In many cases, however, we are interested in what would happen if two or more pieces of input data are changed simultaneously. The easiest way to examine the effect of simultaneous changes is to change the parameters in the model and re-solve the model.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.16 The Excel Sensitivity Report for the MHT Banner Ad Problem

Variable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
$B$20	Business Nile	550000	0	0.00465	0.0003	0.0003
$C$20	Business Zstar	2E+06	0	0.00795	1E+30	0.0003
$D$20	Business Cheetah	0	-0.0014	0.003	0.00135	1E+30
$E$20	Business Stride	0	-0.0026	0.0051	0.00255	1E+30
$F$20	Business Stove	0	-0.0003	0.00315	0.0003	1E+30
$B$21	Science Nile	950000	0	0.00495	0.0003	0.0003
$C$21	Science Zstar	0	-0.005	0.0033	0.00495	1E+30
$D$21	Science Cheetah	0	-0.0009	0.00375	0.0009	1E+30
$E$21	Science Stride	1E+06	0	0.00795	0.0003	0.00225
$F$21	Science Stove	100000	0	0.00375	0.00165	0.0003
$B$22	Sports Nile	0	-0.0017	0.00435	0.00165	1E+30
$C$22	Sports Zstar	0	-0.0022	0.00705	0.00225	1E+30
$D$22	Sports Cheetah	1E+06	0	0.0057	0.0033	0.0009
$E$22	Sports Stride	0	-0.0022	0.00675	0.00225	1E+30
$F$22	Sports Stove	1E+06	0	0.0048	0.0009	0.00165

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$B$23	Nile	2E+06	0	2000000	1E+30	500000
$C$23	Zstar	2E+06	0.0003	2000000	100000	900000
$D$23	Cheetah	1E+06	0	1800000	1E+30	700000
$E$23	Stride	1E+06	0	2000000	1E+30	900000
$F$23	Stove	2E+06	0	2000000	1E+30	500000
$B$23	Nile	2E+06	-0.003	1500000	100000	900000
$C$23	Zstar	2E+06	0	1000000	1000000	1E+30
$D$23	Cheetah	1E+06	-0.0033	1100000	100000	100000
$E$23	Stride	1E+06	0	1000000	100000	1E+30
$F$23	Stove	2E+06	-0.0042	1500000	100000	100000
$G$20	Business Impressions Available	3E+06	0.00765	2550000	900000	100000
$G$21	Science Impressions Available	2E+06	0.00795	2150000	900000	100000
$G$22	Sports Impressions Available	3E+06	0.009	2500000	100000	100000

Assigning Project Leaders to Clients

A linear program known as the assignment problem arises in a variety of decision-making situations. The general form of an **assignment problem** involves assigning "agents" to "tasks." These agent-to-task assignments can involve assigning jobs to machines, sales personnel to sales territories, contracts to bidders, and so on. A distinguishing feature of the assignment problem is that one agent is assigned to at most one task. Specifically, we look for the set of assignments that will optimize a stated objective, such as minimize cost, minimize time required, or maximize profits.

To illustrate the assignment problem, let us consider the case of Fowle Marketing Research, which has just received requests for market research studies from three new clients. The company faces the task of assigning a project leader (agent) to each client (task). Currently, three individuals have no other commitments and are available for the project leader assignments. Fowle's management realizes, however, that the time required to complete each study will depend on the experience and ability of the project leader assigned. The three

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

projects have approximately the same priority, and management wants to assign project leaders to minimize the total number of days required to complete all three projects. If a project leader is to be assigned to one client only, which assignments should be made?

To answer the assignment question, Fowle's management must first consider all possible project leader–client assignments and then estimate the corresponding project completion times. With three project leaders and three clients, nine assignment alternatives are possible. The alternatives and the estimated project completion times in days are summarized in Table 14.6.

Figure 14.17 shows the network representation of Fowle's assignment problem. The nodes correspond to the project leaders and clients, and the arcs represent the possible assignments of project leaders to clients. The supply at each origin node and the demand at each destination node are 1; the cost of assigning a project leader to a client is the time it takes that project leader to complete the client's task. Note the similarity between the network models of the assignment problem (Figure 14.17) and the transportation problem (Figure 14.13). The assignment problem is a special case of the transportation problem in which all supply and demand values equal 1, and the amount shipped over each arc is either 0 or 1. This means that when we solve the assignment problem linear program the optimal values of the decision variables will be 0 or 1, not fractional.

Because the assignment problem is a special case of the transportation problem, we can use the linear programming formulation for the transportation problem to solve the assignment problem. Again, we need a constraint for each node and a variable for each arc. Recall that in the transportation problem the double-subscripted decision variables x_{ij} denoted the number of units shipped from node i to node j. Because the assignment problem has the special structure shown in Figure 14.17 (like the transportation problem), the value of each x_{ij} variable will be integer and be 0 or 1. Therefore, if $x_{11} = 1$, we interpret this as "project leader 1 (Terry) is assigned to client 1," or if $x_{11} = 0$, we interpret this as "project leader 1 (Terry) is not assigned to client 1." In general, we interpret the decision variables for Fowle's assignment problem as

$$x_{ij} = \begin{cases} 1 & \text{if project leader } i \text{ is assigned to client } j \\ 0 & \text{otherwise} \end{cases}$$

where $i = 1, 2, 3$, and $j = 1, 2, 3$

Using this notation and the completion time data in Table 14.6, we develop completion time expressions:

$$\begin{aligned} &\text{Days required for Terry's assignment} && = 10x_{11} + 15x_{12} + 9x_{13} \\ &\text{Days required for Carle's assignment} && = 9x_{21} + 18x_{22} + 5x_{23} \\ &\text{Days required for McClymonds's assignment} && = 6x_{31} + 14x_{32} + 3x_{33} \end{aligned}$$

The sum of the completion times for the three project leaders will provide the total days required to complete the three assignments. Thus, the objective function is

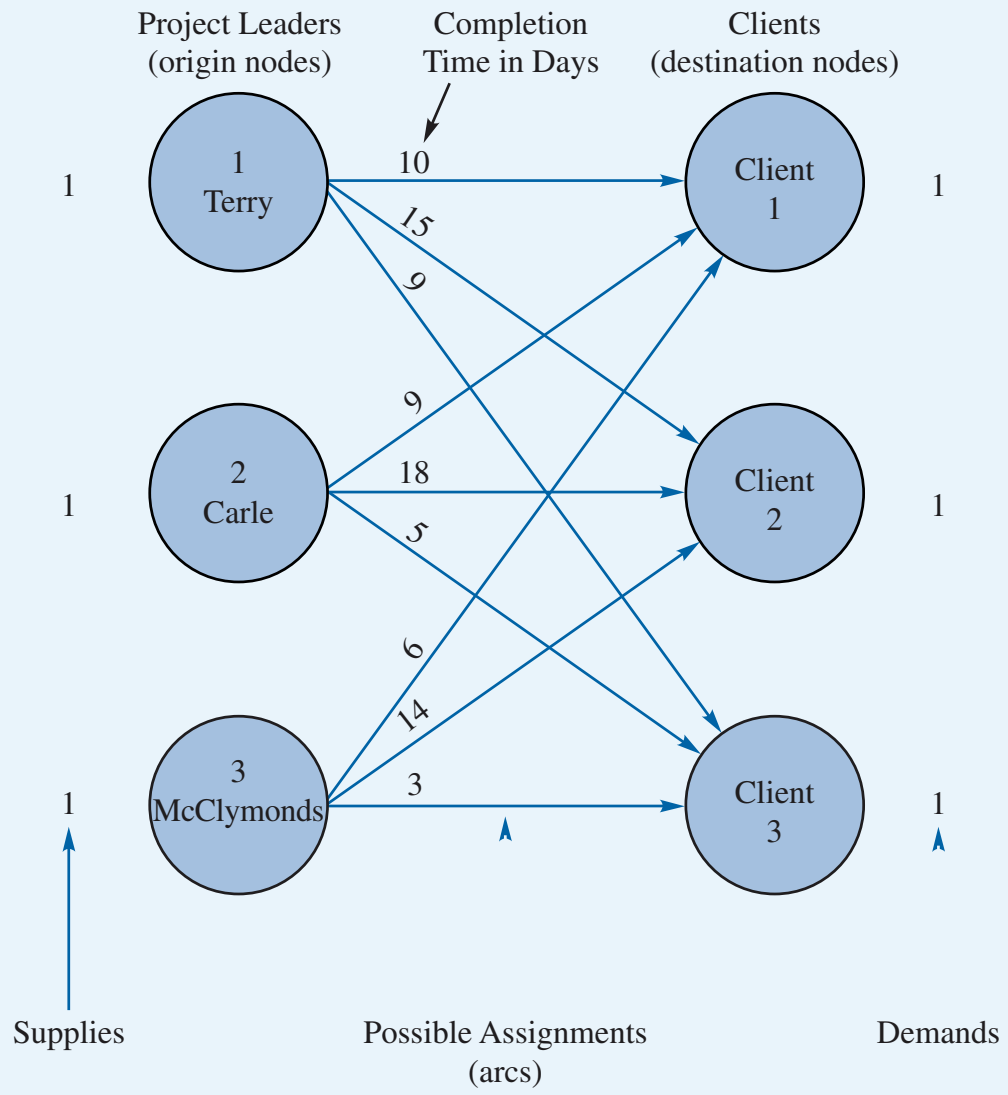
$$\text{Min } 10x_{11} + 15x_{12} + 9x_{13} + 9x_{21} + 18x_{22} + 5x_{23} + 6x_{31} + 14x_{32} + 3x_{33}$$

Table 14.6 Estimated Project Completion Times (Days) for the Fowle Marketing Research Assignment Problem

	Client		
Project Leader	**1**	**2**	**3**
1. Terry	10	15	9
2. Carle	9	18	5
3. McClymonds	6	11	3

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.17 A Network Model of the Fowle Marketing Research Assignment Problem



Because the number of project leaders equals the number of clients, all the constraints could be written as equalities. But when the number of project leaders exceeds the number of clients, less-than-or-equal-to constraints must be used for the project leader constraints.

The constraints for the assignment problem reflect the conditions that each project leader can be assigned to at most one client and that each client must have one assigned project leader. These constraints are written as follows:

$$x_{11} + x_{12} + x_{13} \leq 1 \quad \text{Terry's assignment}$$
$$x_{21} + x_{22} + x_{23} \leq 1 \quad \text{Carle's assignment}$$
$$x_{31} + x_{32} + x_{33} \leq 1 \quad \text{McClymonds' assignment}$$
$$x_{11} + x_{21} + x_{31} = 1 \quad \text{Client 1}$$
$$x_{12} + x_{22} + x_{32} = 1 \quad \text{Client 2}$$
$$x_{13} + x_{23} + x_{33} = 1 \quad \text{Client 3}$$

Note that each node in Figure 14.17 has one constraint. Combining the objective function and constraints into one model provides the following linear programming model of the Fowle Marketing Research assignment problem:

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$$\text{Min } 10x_{11} + 15x_{12} + 9x_{13} + 9x_{21} + 18x_{22} + 5x_{23} + 6x_{31} + 14x_{32} + 3x_{33}$$

s.t.

$$\begin{aligned}
x_{11} + x_{12} + x_{13} &\leq 1\\
x_{21} + x_{22} + x_{23} &\leq 1\\
x_{31} + x_{32} + x_{33} &\leq 1\\
x_{11} + x_{21} + x_{31} &= 1\\
x_{12} + x_{22} + x_{32} &= 1\\
x_{13} + x_{23} + x_{33} &= 1\\
x_{ij} &\geq 0 \quad \text{for } i = 1, 2, 3 \text{ and } j = 1, 2, 3
\end{aligned}$$

Figure 14.18 shows the optimal solution for this model. The same solution produced using Excel Solver is contained in the file *fowle*. Terry is assigned to client 2 ($x_{12} = 1$), Carle is assigned to client 3 ($x_{23} = 1$), and McClymonds is assigned to client 1 ($x_{31} = 1$). The optimal total completion time required is 26 days.

Diet Planning

One of the earliest proposed linear programming problems involves planning a diet. The **diet problem** is to find a cost-minimizing diet that satisfies nutritional requirements. Applications of this basic problem are numerous in agriculture.

Figure 14.18 Spreadsheet Model and Solution for the Fowle Marketing Research Assignment Problem

	A	B	C	D	E	F
1	**Fowle Marketing Research**					
2	**Parameters**					
3			**Client**			
4	**Project Leader**	1	2	3		
5	Terry	10	15	9		
6	Carle	9	18	5		
7	McClymonds	6	14	3		
8						
9	**Model**					
10						
11	**Total Completion Time**	=SUMPRODUCT(B5:D7,B15:D17)				
12						
13						
14	**Project Leader**	to Client 1	to Client 2	to Client 3	Total	
15	Terry	0	1	0	=SUM(B15:D15)	
16	Carle	0	0	1	=SUM(B16:D16)	
17	McClymonds	1	0	0	=SUM(B17:D17)	
18	Total	=SUM(B15:B17)	=SUM(C15:C17)	=SUM(D15:D17)		
19						

MODEL *file*
fowle

	A	B	C	D	E	F
1	**Fowle Marketing Research**					
2	**Parameters**					
3			**Client**			
4	**Project Leader**	1	2	3		
5	Terry	10	15	9		
6	Carle	9	18	5		
7	McClymonds	6	14	3		
8						
9	**Model**					
10						
11	**Total Completion Time**	26				
12						
13						
14	**Project Leader**	to Client 1	to Client 2	to Client 3	Total	
15	Terry	0	1	0	1	
16	Carle	0	0	1	1	
17	McClymonds	1	0	0	1	
18	Total	1	1	1		
19						

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 14.7 Data for the Springer Farms Dairy Cow Diet Problem

	Shelled Corn	Soybean Meal	Dicalcium Phosphate	Limestone	Alfalfa	Corn Silage	Minimum Daily	Maximum Daily
Cost ($/kg)	0.11	0.24	0.33	0.08	0.07	0.06		
Dry Matter (kg/kg)	1.00	1.00	1.00	1.00	1.00	1.00	0.00	19.63
Molar Mass (Mcal/kg)	2.09	1.85	0.00	0.00	1.21	1.54	29.10	
Protein (kg/kg)	0.10	0.50	0.00	0.00	0.18	0.08	2.45	
Calcium (g/kg)	20.00	3.60	210.00	380.00	12.40	2.80	97.00	
Phosphorous (g/kg)	2.60	7.50	180.00	0.00	2.30	2.20	69.00	
Alfalfa (kg/kg)	0.00	0.00	0.00	0.00	1.00	0.00	3.00	

Let us consider the case of Springer Farms. Springer has 11 dairy farms in the upper Midwest. Sarah Springer, the owner of Springer Farms, must ensure the dairy cows are healthy and productive while controlling the cost of feeding the cows. By determining a cost-effective diet that meets a single cow's daily nutritional requirements, the solution can be scaled to the overall herd by simply multiplying the solution by the number of cows. So, the diet problem to solve is to minimize the cost of feeding one cow for one day subject to satisfying the daily nutritional requirements of the cow. Table 14.7 contains the basic data we need to solve this problem.[2]

The mix of feed for the cow can contain a mixture of six ingredients (shelled corn, soybean meal, dicalcium phosphate, limestone, alfalfa, and corn silage). The first row of data in Table 14.7 gives the cost per kilogram for each of these six ingredients. There are six nutritional considerations that involve the amount of dry matter, molar mass (the mass of one mole of a substance measured in Mcal = 1,000 calories), protein, calcium, phosphorous and alfalfa. The table gives the contribution per kilogram of each ingredient of the particular nutritional factor. For example, a kilogram of soybean meal contributes 0.50 kilogram of protein and 3.60 grams of calcium. No more than 19.63 kilograms of dry matter can be fed to the cow per day as shown in the Maximum Daily column. The other five requirements are minimum daily requirements. For example, at least 97 grams of calcium must be in the daily mix. The other minimum daily requirements are similar.

The decision to be made is how much of each of the six ingredients (in kilograms) to include in the mix for a cow per day. We define the following decision variables:

x_i = amount (in kilograms) of ingredient i to include in the mix for a cow per day
$i = 1, 2, 3, 4, 5, 6$

where $i = 1$ refers to shelled corn, $i = 2$ refers to soybean meal, $i = 3$ refers to dicalcium phosphate, $i = 4$ refers to limestone, $i = 5$ refers to alfalfa, and $i = 6$ refers to corn silage. The objective function is to minimize the daily cost of feeding a cow:

$$\text{Min } 0.11x_1 + 0.24x_2 + 0.33x_3 + 0.08x_4 + 0.07x_5 + 0.06x_6$$

The first nutritional constraint is that the total amount of dry matter fed to a cow per day cannot exceed 19.63 kilograms:

$$1.00x_1 + 1.00x_2 + 1.00x_3 + 1.00x_4 + 1.00x_5 + 1.00x_6 \leq 19.63$$

The other five constraints enforce minimum daily requirements. For example, the Molar Mass constraint is as follows:

$$2.09x_1 + 1.85x_2 + 0.00x_3 + 0.00x_4 + 1.21x_5 + 1.54x_6 \geq 29.10$$

The other four constraints are similar. The model combining the objective function and constraints and including nonnegativity constraints is as follows:

[2] The data in Table 14.7 are from Black. J.R and J. Hlubik, "Symposium: Computer Programs for Dairy Cattle Feeding and Management—Past, Present, and Future," *Journal of Dairy Science*, 63 (1980):1366–1378.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 14.19 Spreadsheet Model and Solution for the Springer Farms Cattle Feed Problem

	A	B	C	D	E	F	G	H	I	J
1	**Springer Dairy Farms**									
2										
3	**Parameters**									
4		Shelled	Soybean	Dicalcium			Corn			
5		Corn	Meal	Phosphate	Limestone	Alfalfa	Silage			
6	Cost ($/kg)	0.11	0.24	0.33	0.08	0.066	0.063	Minimum Daily	Maximum Daily	Units
7	Dry Matter (kg)	1	1	1	1	1	1	0	19.63	kg
8	Molar Mass (Mcal/kg)	2.09	1.85	0	0	1.21	1.54	29.1		Mcal
9	Protein (kg/kg)	0.1	0.5	0	0	0.18	0.08	2.45		kg
10	Calcium (g/kg)	20	3.6	210	380	12.4	2.8	97		grams
11	Phosphorous (g/kg)	2.6	7.5	180	0	2.3	2.2	69		grams
12	Alfalfa (kg/kg)	0	0	0	0	1	0	3		kg
13										
14	**Model**									
15		Shelled	Soybean	Dicalcium			Corn			
16		Corn	Meal	Phosphate	Limestone	Alfalfa	Silage			
17	Amount (kg)	0	13.7675675675676	2.86243243243244	0	3	0			
18										
19	Cost/Day/Cow ($)	=SUMPRODUCT(B6:G6,B17:G17)								
20										
21		Total	Minimum Daily	Maximum Daily						
22	Dry Matter (kg)	=SUMPRODUCT(B7:G7,$B$17:$G$17)	=H7	=I7						
23	Molar Mass (Mcal)	=SUMPRODUCT(B8:G8,$B$17:$G$17)	=H8							
24	Protein (kg)	=SUMPRODUCT(B9:G9,$B$17:$G$17)	=H9							
25	Calcium (g)	=SUMPRODUCT(B10:G10,$B$17:$G$17)	=H10							
26	Phosphorous (g)	=SUMPRODUCT(B11:G11,$B$17:$G$17)	=H11							
27	Alfalfa (kg)	=SUMPRODUCT(B12:G12,$B$17:$G$17)	=H12							

MODEL *file*
cattlefeed

	A	B	C	D	E	F	G	H	I	J
1	**Springer Dairy Farms**									
2										
3	**Parameters**									
4		Shelled	Soybean	Dicalcium			Corn			
5		Corn	Meal	Phosphate	Limestone	Alfalfa	Silage			
6	Cost ($/kg)	0.11	0.24	0.33	0.08	0.07	0.06	Minimum Daily	Maximum Daily	Units
7	Dry Matter (kg)	1.00	1.00	1.00	1.00	1.00	1.00	0.00	19.63	kg
8	Molar Mass (Mcal/kg)	2.09	1.85	0.00	0.00	1.21	1.54	29.10		Mcal
9	Protein (kg/kg)	0.10	0.50	0.00	0.00	0.18	0.08	2.45		kg
10	Calcium (g/kg)	20.00	3.60	210.00	380.00	12.40	2.80	97.00		grams
11	Phosphorous (g/kg)	2.60	7.50	180.00	0.00	2.30	2.20	69.00		grams
12	Alfalfa (kg/kg)	0.00	0.00	0.00	0.00	1.00	0.00	3.00		kg
13										
14	**Model**									
15		Shelled	Soybean	Dicalcium			Corn			
16		Corn	Meal	Phosphate	Limestone	Alfalfa	Silage			
17	Amount (kg)	0.00	13.77	2.86	0.00	3.00	0.00			
18										
19	Cost/Day/Cow ($)	$4.45								
20										
21		Total	Minimum Daily	Maximum Daily						
22	Dry Matter (kg)	19.63	0.00	19.63						
23	Molar Mass (Mcal)	29.10	29.10							
24	Protein (kg)	7.42	2.45							
25	Calcium (g)	687.87	97.00							
26	Phosphorous (g)	625.39	69.00							
27	Alfalfa (kg)	3.00	3.00							
28										

$$
\begin{aligned}
\text{Min } & 0.11x_1 + 0.24x_2 + 0.33x_3 + 0.08x_4 + 0.07x_5 + 0.06x_6 \\
\text{s.t.} \\
& 1.00x_1 + 1.00x_2 + 1.00x_3 + 1.00x_4 + 1.00x_5 + 1.00x_6 \geq 19.63 \quad \text{(Dry Matter)} \\
& 2.09x_1 + 1.85x_2 + 1.21x_5 + 1.54x_6 \geq 29.10 \quad \text{(Molar Mass)} \\
& 0.10x_1 + 0.50x_2 + 0.18x_5 + 0.08x_6 \geq 2.45 \quad \text{(Protein)} \\
& 20.00x_1 + 3.60x_2 + 210.00x_3 + 380.00x_4 + 12.40x_5 + 2.80x_6 \geq 97.00 \quad \text{(Calcium)} \\
& 2.60x_1 + 7.50x_2 + 180.00x_3 + 2.30x_5 + 2.20x_6 \geq 69.00 \quad \text{(Phosphorous)} \\
& 1.00x_5 \geq 3.00 \quad \text{(Alfalfa)} \\
& x_1, x_2, x_3, x_4, x_5, x_6 \geq 0
\end{aligned}
$$

A spreadsheet model and solution are shown in Figure 14.19. The optimal objective function value is $4.45 per cow per day by using 13.77 kilograms of soybean meal, 2.86 kilograms of dicalcium phosphate and 3.00 kilograms of alfalfa. The optimal diet includes no shelled corn, limestone or corn silage.

Notes & Comments

Network models such as the transportation problem and the assignment problem will yield optimal variable values that are integer when the right-hand-side data are integer. This is very important for the assignment problem, since the model is dependent on the 0 or 1 defintion of the variables.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

14.7 Generating an Alternative Optimal Solution for a Linear Program

The goal of business analytics is to provide information to management for improved decision making. If a linear program has more than one optimal solution, as discussed in Section 14.4, it would be good for management to know this. There might be factors external to the model that make one optimal solution preferable to another. For example, in a portfolio optimization problem, perhaps more than one strategy yields the maximum expected return. However, those strategies might be quite different in terms of their risk to the investor. Knowing the optimal alternatives and then assessing the risk of each, the investor could then pick the least risky alternative from the optimal solutions. In this section, we discuss how to generate an alternative optimal solution if one exists.

Let us reconsider the Foster Generators transportation problem from the previous section. If one exists, how might we generate an alternative optimal solution for this problem? From Figure 14.14 we know that the following is an optimal solution:

$$x_{11} = 1{,}000, x_{12} = 4{,}000, x_{13} = 0, x_{14} = 0$$
$$x_{21} = 2{,}500, x_{22} = 0, x_{23} = 2{,}000, x_{24} = 1{,}500$$
$$x_{31} = 2{,}500, x_{32} = 0, x_{33} = 0, x_{34} = 0$$

The optimal cost is $39,500. With this information, we may revise our previous model to try to find an alternative optimal solution. We know that any alternative solution must be feasible, so it must satisfy all of the constraints of the original model. Also, to be optimal, the solution must give a total cost of $39,500. We can enforce this by taking the objective function and making it a constraint equal to $39,500:

$$3x_{11} + 2x_{12} + 7x_{13} + 6x_{14} + 6x_{21} + 5x_{22} + 2x_{23} + 3x_{24} + 2x_{31} + 5x_{32} + 4x_{33} + 5x_{34} = 39{,}500.$$

But, what should our objective function be for the revised problem? In the solution we previously found:

$$x_{13} = x_{14} = x_{22} = x_{32} = x_{33} = x_{34} = 0$$

If we maximize the sum of these variables and if the optimal objective function value of this revised problem is positive, we have found a different feasible solution that is also optimal. The revised model is as follows:

$$\begin{aligned}
\text{Max} \quad & x_{13} + x_{14} + x_{22} + x_{32} + x_{33} + x_{34} \\
\text{s.t.} \quad & \\
& x_{11} + x_{12} + x_{13} + x_{14} \le 5{,}000 \\
& x_{21} + x_{22} + x_{23} + x_{24} \le 6{,}000 \\
& x_{31} + x_{32} + x_{33} + x_{34} \le 2{,}500 \\
& x_{11} + x_{21} + x_{31} = 6{,}000 \\
& x_{12} + x_{22} + x_{32} = 4{,}000 \\
& x_{13} + x_{23} + x_{33} = 2{,}000 \\
& x_{14} + x_{24} + x_{34} = 1{,}500 \\
& 3x_{11} + 2x_{12} + 7x_{13} + 6x_{14} + 6x_{21} + 5x_{22} + 2x_{23} + 3x_{24} + 2x_{31} + 5x_{32} + 4x_{33} + 5x_{34} = 39{,}500 \\
& x_{ij} \ge 0 \\
& \text{for } i = 1, 2, 3 \text{ and } j = 1, 2, 3, 4
\end{aligned}$$

The solution to this problem has an objective function value of 2,500, indicating that the variables that were zero in the previous solution now add up to 2,500. The new solution is shown in Table 14.8.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 14.8 An Alternative Optimal Solution to the Foster Generators Transportation Problem

Total Cost = $39,500

		Amount Shipped To:				
		Boston	Chicago	St. Louis	Lexington	**Total**
From:	Cleveland	3,500	1,500	0	0	5,000
	Bedford	0	2,500	2,000	1,500	6,000
	York	2,500	0	0	0	2,500
	Total	6,000	4,000	2,000	1,500	

Comparing Figure 14.14 and Table 14.8, we see that in this new solution, Bedford ships 2,500 units to Chicago instead of to Boston.

What types of issues might make management prefer one of these solutions over the other? Notice that the original solution has the Boston distribution center sourced from all three plants, whereas each of the other distribution centers is sourced by one plant. This would imply that the manager in the Boston distribution center has to deal with three different plant managers, whereas each of the other distribution center managers has only one plant manager. The Boston manager might feel disadvantaged, having to spend too much time coordinating among the plants. The alternative solution provides a more balanced solution. Managers in Boston and Chicago each deal with two plants, and those in St. Louis and Lexington, which have lower total volumes, deal with only one plant. Because the alternative solution seems to be more equitable, it might be preferred. Recall that both solutions give a total cost of $39,500.

In summary, the general approach for trying to find an alternative optimal solution to a linear program is as follows:

Step 1. Solve the linear program

Step 2. Make a new objective function to be maximized. It is the sum of those variables that were equal to zero in the solution from Step 1

Step 3. Keep all the constraints from the original problem. Add a constraint that forces the original objective function to be equal to the optimal objective function value from Step 1

Step 4. Solve the problem created in Steps 2 and 3. If the objective function value is positive, you have found an alternative optimal solution

Notes & Comments

Steps 1–4 for finding an alternative optimal solution may be repeated to try to find more than one alternative optimal solution. However, the process is not guaranteed to find an alternative optimal solution when one exists. For example, alternative optimal solutions that are not an extreme point (see Figure 14.8) will not be found by this approach.

Summary

We formulated linear programming models for the Par, Inc. maximization problem and the M&D Chemicals minimization problem. For the Par, Inc. problem, we showed how a graphical solution procedure could be used to solve a two-variable problem to help us better understand how the computer can solve large linear programs. We discussed how Excel Solver can

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

be used to solve linear optimization problems. In formulating a linear programming model of the Par, Inc. and M&D problems, we developed a general definition of a linear program.

A linear program is a mathematical model with the following qualities:

1. A linear objective function that is to be maximized or minimized
2. A set of linear constraints
3. Variables restricted to nonnegative values

Slack variables may be used to write less-than-or-equal-to constraints in equality form, and surplus variables may be used to write greater-than-or-equal-to constraints in equality form. The value of a slack variable can usually be interpreted as the amount of unused resource, whereas the value of a surplus variable indicates the amount over and above some stated minimum requirement. Binding constraints have zero slack or surplus.

If the solution to a linear program is infeasible or unbounded, no optimal solution to the problem can be found. In the case of infeasibility, no feasible solutions are possible. In the case of an unbounded solution, the objective function can be made infinitely large for a maximization problem and infinitely small for a minimization problem. In the case of alternative optimal solutions, two or more optimal extreme points exist.

We also discussed sensitivity analysis and the interpretation of the Sensitivity Report generated by Excel Solver and how the impact of changes in the objective function coefficients and right-hand side values of constraints can be assessed. We showed how to write a mathematical model using general linear programming notation and presented additional examples of linear programming applications: portfolio selection, transportation planning, media selection, the diet problem, and the assignment problem. Finally, we concluded the chapter with a procedure for finding an alternative optimal solution when one exists.

Glossary

Alternative optimal solutions The case in which more than one solution provides the optimal value for the objective function.

Assignment problem An optimization problem that involves the assignment of agents to tasks where an agent is assigned to at most one task and every task is assigned exactly one agent.

Binding constraint A constraint that holds as an equality at the optimal solution.

Constraints Restrictions that limit the settings of the decision variables.

Decision variable A controllable input for a linear programming model.

Diet problem An optimization problem for which the objective is to find a cost-minimizing diet to satisfy nutritional requirement constraints.

Extreme point Graphically speaking, the feasible solution points occurring at the vertices, or "corners," of the feasible region. With two-variable problems, extreme points are determined by the intersection of the constraint lines.

Feasible region The set of all feasible solutions.

Feasible solution A solution that satisfies all the constraints simultaneously.

Infeasibility The situation in which no solution to the linear programming problem satisfies all the constraints.

Linear function A mathematical function in which each variable appears in a separate term and is raised to the first power.

Linear programming model (linear program) A mathematical model with a linear objective function, a set of linear constraints, and nonnegative variables.

Mathematical model A representation of a problem in which the objective and all constraint conditions are described by mathematical expressions.

Nonnegativity constraints A set of constraints that requires all variables to be nonnegative.

Objective function The expression that defines the quantity to be maximized or minimized in a linear programming model.

Objective function coefficient allowable increase (decrease) The allowable increase/decrease of an objective function coefficient is the amount the coefficient may increase (decrease) without causing any change in the values of the decision variables in the optimal

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

solution. The allowable increase/decrease for the objective function coefficients can be used to calculate the range of optimality.

Problem formulation (modeling) The process of translating a verbal statement of a problem into a mathematical statement called the *mathematical model*.

Reduced cost If a variable is at its lower bound of zero, the reduced cost is equal to the shadow price of the nonnegativity constraint for that variable. In general, if a variable is at its lower or upper bound, the reduced cost is the shadow price for that simple lower- or upper-bound constraint.

Right-hand side allowable increase (decrease) The amount the right-hand side may increase (decrease) without causing any change in the shadow price for that constraint. The allowable increase and decrease for the right-hand side can be used to calculate the range of feasibility for that constraint.

Sensitivity analysis The study of how changes in the input parameters of a linear programming problem affect the optimal solution.

Shadow price The change in the optimal objective function value per unit increase in the right-hand side of a constraint.

Slack The difference between the right-hand-side and the left-hand-side of a less-than-or-equal-to constraint.

Slack variable A variable added to the left-hand side of a less-than-or-equal-to constraint to convert the constraint into an equality. The value of this variable can usually be interpreted as the amount of unused resources.

Surplus variable A variable subtracted from the left-hand side of a greater-than-or-equal-to constraint to convert the constraint into an equality. The value of this variable can usually be interpreted as the amount over and above some required minimum level.

Transportation problem An optimization problem that often involves minimizing the cost of shipping goods from a set of origin locations to a set of destination locations.

Unbounded The situation in which the value of the solution may be made infinitely large in a maximization linear programming problem or infinitely small in a minimization problem without violating any of the constraints.

Problems

1. **Sporting Goods Product Mix.** Kelson Sporting Equipment, Inc. makes two types of baseball gloves: a regular model and a catcher's model. The firm has 900 hours of production time available in its cutting and sewing department, 300 hours available in its finishing department, and 100 hours available in its packaging and shipping department. The production time requirements and the profit contribution per glove are given in the following table:

DATA*file*
sportinggoods

	Production Time (hours)			
Model	Cutting and Sewing	Finishing	Packaging and Shipping	Profit/Glove
Regular model	1	½	⅛	$5
Catcher's model	½	⅓	¼	$8

Assuming that the company is interested in maximizing the total profit contribution, answer the following questions. **LO 1, 2, 3, 4**

a. What is the linear programming model for this problem?
b. Develop a spreadsheet model and find the optimal solution using Excel Solver. How many of each model should Kelson manufacture?
c. What is the total profit contribution Kelson can earn with the optimal production quantities?
d. How many hours of production time will be scheduled in each department?
e. What is the slack time in each department?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

2. **Advertising Budget Allocation.** The Sea Wharf Restaurant would like to determine the best way to allocate a monthly advertising budget of $1,000 between newspaper advertising and radio advertising. Management decided that at least 25% of the budget must be spent on each type of media and that the amount of money spent on local newspaper advertising must be at least twice the amount spent on radio advertising. A marketing consultant developed an index that measures audience exposure per dollar of advertising on a scale from 0 to 100, with higher values implying greater audience exposure. If the value of the index for local newspaper advertising is 50 and the value of the index for spot radio advertising is 80, how should the restaurant allocate its advertising budget to maximize the value of total audience exposure? **LO 1, 2**
   a. Formulate a linear programming model that can be used to determine how the restaurant should allocate its advertising budget in order to maximize the value of total audience exposure.
   b. Develop a spreadsheet model and solve the problem using Excel Solver.
3. **Building a Financial Portfolio.** Blair & Rosen, Inc. (B&R) is a brokerage firm that specializes in investment portfolios designed to meet the specific risk tolerances of its clients. A client who contacted B&R this past week has a maximum of $50,000 to invest. B&R's investment advisor decides to recommend a portfolio consisting of two investment funds: an Internet fund and a Blue Chip fund. The Internet fund has a projected annual return of 12%, and the Blue Chip fund has a projected annual return of 9%. The investment advisor requires that at most $35,000 of the client's funds should be invested in the Internet fund. B&R services include a risk rating for each investment alternative. The Internet fund, which is the more risky of the two investment alternatives, has a risk rating of 6 per $1,000 invested. The Blue Chip fund has a risk rating of 4 per $1,000 invested. For example, if $10,000 is invested in each of the two investment funds, B&R's risk rating for the portfolio would be $6(10) + 4(10) = 100$. Finally, B&R developed a questionnaire to measure each client's risk tolerance. Based on the responses, each client is classified as a conservative, moderate, or aggressive investor. Suppose that the questionnaire results classified the current client as a moderate investor. B&R recommends that a client who is a moderate investor limit his or her portfolio to a maximum risk rating of 240. **LO 1, 2, 3**
   a. Formulate a linear programming model to find the best investment strategy for this client.
   b. Build a spreadsheet model and solve the problem using Excel Solver. What is the recommended investment portfolio for this client? What is the annual return for the portfolio?
   c. Suppose that a second client with $50,000 to invest has been classified as an aggressive investor. B&R recommends that the maximum portfolio risk rating for an aggressive investor is 320. What is the recommended investment portfolio for this aggressive investor?
   d. Suppose that a third client with $50,000 to invest has been classified as a conservative investor. B&R recommends that the maximum portfolio risk rating for a conservative investor is 160. Develop the recommended investment portfolio for the conservative investor.
4. **Bank Loan Funds Allocation.** Adirondack Savings Bank (ASB) has $1 million in new funds that must be allocated to home loans, personal loans, and automobile loans. The annual rates of return for the three types of loans are 7% for home loans, 12% for personal loans, and 9% for automobile loans. The bank's planning committee has decided that at least 40% of the new funds must be allocated to home loans. In addition, the planning committee has specified that the amount allocated to personal loans cannot exceed 60% of the amount allocated to automobile loans. **LO 1, 2, 3, 5, 6, 7**
   a. Formulate a linear programming model that can be used to determine the amount of funds ASB should allocate to each type of loan to maximize the total annual return for the new funds.
   b. How much should be allocated to each type of loan? What is the total annual return? What is the annual percentage return?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. If the interest rate on home loans increases to 9%, would the amount allocated to each type of loan change? Explain.
d. Suppose the total amount of new funds available is increased by $10,000. What effect would this have on the total annual return? Explain.
e. Assume that ASB has the original $1 million in new funds available and that the planning committee has agreed to relax the requirement that at least 40% of the new funds must be allocated to home loans by 1%. How much would the annual return change? How much would the annual percentage return change?

5. **Hotel Room Reservations.** Round Tree Manor is a hotel that provides two types of rooms with three rental classes: Super Saver, Deluxe, and Business. The profit per night for each type of room and rental class is as follows:

DATA*file* hotelrooms

	Rental Class		
Room	**Super Saver**	**Deluxe**	**Business**
Type I (Mountain View)	$30	$35	—
Type II (Street View)	$20	$30	$40

Round Tree's management makes a forecast of the demand by rental class for each night in the future. A linear programming model developed to maximize profit is used to determine how many reservations to accept for each rental class. The demand forecast for a particular night is 130 rentals in the Super Saver class, 60 in the Deluxe class, and 50 in the Business class. Since these are the forecasted demands, Round Tree will take no more than these amounts of each reservation for each rental class. Round Tree has a limited number of each type of room. There are 100 Type I rooms and 120 Type II rooms. **LO 1, 2, 3, 4, 6**

a. Formulate and solve a linear program to determine how many reservations to accept in each rental class and how the reservations should be allocated to room types.
b. For the solution in part (a), how many reservations can be accommodated in each rental class? Is the demand for any rental class not satisfied?
c. With a little work, an unused office area could be converted to a rental room. If the conversion cost is the same for both types of rooms, would you recommend converting the office to a Type I or a Type II room? Why?
d. Could the linear programming model be modified to plan for the allocation of rental demand for the next night? What information would be needed and how would the model change?

6. **Labor Allocation in a Design Project.** Industrial Designs has been awarded a contract to design a label for a new wine produced by Lake View Winery. The company estimates that 150 hours will be required to complete the project. The firm's three graphic designers available for assignment to this project are Lisa, a senior designer and team leader; David, a senior designer; and Sarah, a junior designer. Because Lisa has worked on several projects for Lake View Winery, management specified that Lisa must be assigned at least 40% of the total number of hours assigned to the two senior designers. To provide label designing experience for Sarah, the junior designer must be assigned at least 15% of the total project time. However, the number of hours assigned to Sarah must not exceed 25% of the total number of hours assigned to the two senior designers. Due to other project commitments, Lisa has a maximum of 50 hours available to work on this project. Hourly wage rates are $30 for Lisa, $25 for David, and $18 for Sarah. **LO 1, 2, 3, 6, 7**

a. Formulate a linear program that can be used to determine the number of hours each graphic designer should be assigned to the project to minimize total cost.
b. How many hours should be assigned to each graphic designer? What is the total cost?
c. Suppose Lisa could be assigned more than 50 hours. What effect would this have on the optimal solution? Explain.
d. If Sarah were not required to work a minimum number of hours on this project, would the optimal solution change? Explain.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

7. **Component Manufacturing.** Vollmer Manufacturing makes three components for sale to refrigeration companies. The components are processed on two machines: a shaper and a grinder. The times (in minutes) required on each machine are as follows:

DATA *file* compmanufacturing

	Machine	
Component	Shaper	Grinder
1	6	4
2	4	5
3	4	2

The shaper is available for 120 hours, and the grinder for 110 hours. No more than 200 units of component 3 can be sold, but up to 1,000 units of each of the other components can be sold. In fact, the company already has orders for 600 units of component 1 that must be satisfied. The per unit selling price and per unit variable cost for each of the three components follows. **LO 1, 2, 3, 5, 6, 7**

Component	Selling Price ($)	Material Cost ($)	Labor Cost ($)
1	25	12	5
2	18	8	4
3	27	13	5

a. For each component, calculate the profit margin (profit margin = selling price − material cost − labor cost). Formulate and solve for the recommended production quantities that will maximize contribution to profit.
b. What are the objective coefficient ranges for the three components? Interpret these ranges for company management.
c. What are the right-hand-side ranges? Interpret these ranges for company management.
d. If more time could be made available on the grinder, how much would it be worth?
e. If more units of component 3 can be sold by reducing the sales price by $4, should the company reduce the price?

8. **Lithium Battery Production.** Photon Technologies, Inc., a manufacturer of batteries for mobile phones, signed a contract with a large electronics manufacturer to produce three models of lithium-ion battery packs for a new line of phones. The contract calls for the following:

Battery Pack	Production Quantity
PT-100	200,000
PT-200	100,000
PT-300	150,000

DATA *file* battery

Photon Technologies can manufacture the battery packs at manufacturing plants located in the Philippines and Mexico. The unit cost of the battery packs differs at the two plants because of differences in production equipment and wage rates. The unit costs for each battery pack at each manufacturing plant are as follows:

	Cost at each Plant	
Product	Philippines ($)	Mexico ($)
PT-100	0.95	0.98
PT-200	0.98	1.06
PT-300	1.34	1.15

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The PT-100 and PT-200 battery packs are produced using similar production equipment available at both plants. However, each plant has a limited capacity for the total number of PT-100 and PT-200 battery packs produced. The combined PT-100 and PT-200 production capacities are 175,000 units at the Philippines plant and 160,000 units at the Mexico plant. The PT-300 production capacities are 75,000 units at the Philippines plant and 100,000 units at the Mexico plant. The cost of shipping from the Philippines plant is $0.18 per unit, and the cost of shipping from the Mexico plant is $0.10 per unit. **LO 1, 2, 3, 5, 6, 7**

a. Develop a linear program that Photon Technologies can use to determine how many units of each battery pack to produce at each plant to minimize the total production and shipping cost associated with the new contract.
b. Solve the linear program developed in part (a) to determine the optimal production plan.
c. Use sensitivity analysis to determine how much the production and/or shipping cost per unit would have to change to produce additional units of the PT-100 in the Philippines plant.
d. Use sensitivity analysis to determine how much the production and/or shipping cost per unit would have to change to produce additional units of the PT-200 in the Mexico plant.

9. **Maximizing a Course Grade.** Cory is an undergraduate student who just finished a course in linear programming. The course grade is composed of the following components: midterm exam score, final exam score, individual assignments score and class participation score. Cory just received their final exam score; so, they know their percentage score on each of the four components of the course grade. Cory earned 86% on the midterm exam, 94% on the final exam, 93% on the individual assignments and 82% on class participation. The default weights for the components are 0.25 each. Hence, Cory's total score is 0.25(86) + 0.25(94) + 0.25(93) + 0.25(82) = 88.75. The total score of 88.75 is a grade of B. A score of 90 and higher will earn an A for the course. Cory's instructor has decided to allow the students to either take their default total score or try to maximize their total score, subject to the following restrictions:

- The participation score weight can be no more than 0.25.
- The midterm exam score weight must be at least twice as much as the participation weight.
- The final exam score weight must be at least three times as much as the participation weight.
- The weight for each of the four components must be at least 0.10.
- The weights used for each component must be nonnegative and the four weights must sum to one.

Cory has decided to try to maximize their grade. **LO 1, 2, 3**

a. Develop a linear programming model that could be used to maximize Cory's total score.
b. Build and solve a spreadsheet model of the formulation from part (a). What are the optimal weights and what is Cory's total score?
c. How much did the model improve Cory's total score. Does Cory's maximized score result in a better course letter grade? Explain.

10. **Production Planning.** The management of Hartman Company is trying to determine the amount of each of two products to produce over the coming planning period. The following information concerns labor availability, labor utilization, and product profitability. **LO 1, 2, 3**

DATA*file*
prodplannning

	Labor-Hours Required (hours/unit)		
Department	Product 1	Product 2	Hours Available
A	1.00	0.35	100
B	0.30	0.20	36
C	0.20	0.50	50
Profit contribution/unit	$30.00	$15.00	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Develop a linear programming model of the Hartman Company problem. Solve the model to determine the optimal production quantities of products 1 and 2.
b. In computing the profit contribution per unit, management does not deduct labor costs because they are considered fixed for the upcoming planning period. However, suppose that overtime can be scheduled in some of the departments. Which departments would you recommend scheduling for overtime? How much would you be willing to pay per hour of overtime in each department?
c. Suppose that 10, 6, and 8 hours of overtime may be scheduled in departments A, B, and C, respectively. The cost per hour of overtime is $18 in department A, $22.50 in department B, and $12 in department C. Formulate a linear programming model that can be used to determine the optimal production quantities if overtime is made available. What are the optimal production quantities, and what is the revised total contribution to profit? How much overtime do you recommend using in each department? What is the increase in the total contribution to profit if overtime is used?

11. **Credit Union Fund Allocation.** The employee credit union at State University is planning the allocation of funds for the coming year. The credit union makes four types of loans to its members. In addition, the credit union invests in risk-free securities to stabilize income. The various revenue-producing investments, together with annual rates of return, are as follows:

DATA*file* creditunion

Type of Loan/Investment	Annual Rate of Return (%)
Automobile loans	8
Furniture loans	10
Other secured loans	11
Signature loans	12
Risk-free securities	9

The credit union will have $2 million available for investment during the coming year. State laws and credit union policies impose the following restrictions on the composition of the loans and investments:

- Risk-free securities may not exceed 30% of the total funds available for investment.
- Signature loans may not exceed 10% of the funds invested in all loans (automobile, furniture, other secured, and signature loans).
- Furniture loans plus other secured loans may not exceed the automobile loans.
- Other secured loans plus signature loans may not exceed the funds invested in risk-free securities.

How should the $2 million be allocated to each of the loan/investment alternatives to maximize total annual return? What is the projected total annual return? **LO 1, 2, 3**

12. **Seafood Buying Strategy.** The Atlantic Seafood Company (ASC) is a buyer and distributor of seafood products that are sold to restaurants and specialty seafood outlets throughout the Northeast. ASC has a frozen storage facility in New York City that serves as the primary distribution point for all products. One of the ASC products is frozen large black tiger shrimp, which are sized at 16 to 20 pieces per pound. Each Saturday, ASC can purchase more tiger shrimp or sell the tiger shrimp at the existing New York City warehouse market price. ASC's goal is to buy tiger shrimp at a low weekly price and sell it later at a higher price. ASC currently has 20,000 pounds of tiger shrimp in storage. Space is available to store a maximum of 100,000 pounds of tiger shrimp each week. In addition, ASC developed the following estimates of tiger shrimp prices for the next four weeks:

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA file seafood

Week	Price/lb
1	$6.00
2	$6.20
3	$6.65
4	$5.55

ASC would like to determine the optimal buying/storing/selling strategy for the next four weeks. The cost to store a pound of shrimp for one week is $0.15, and to account for unforeseen changes in supply or demand, management also indicated that 25,000 pounds of tiger shrimp must be in storage at the end of week 4. Determine the optimal buying/storing/selling strategy for ASC. What is the projected four-week profit? **LO 1, 2, 3**

13. **Umpire Assignments.** The scheduler for a softball league is in the process of scheduling the umpires for each game in the coming week. There are five games in the coming week and six umpires are available. Every game needs exactly one umpire and an umpire can be scheduled for no more than one game in the week. The following table shows the number of times each umpire has been assigned to at least one of the two teams in a game thus far in the season. For example, Umpire 2 has previously been assigned to one of the teams playing in Game 1 three times. The objective of the scheduler is to assign umpires to the games so as to minimize the number of times the umpires have previously been paired with these teams. **LO 1, 2, 3**

Number of times each umpire has been assigned with the teams playing next week:

DATA file umpires

Umpire	Game 1	Game 2	Game 3	Game 4	Game 5
1	2	1	2	2	2
2	3	0	1	4	1
3	2	2	2	2	1
4	3	2	0	2	3
5	3	3	4	1	0
6	0	1	1	5	1

a. Formulate a model that will assign umpires to games and minimize the sum of the number times the umpires have been assigned to their assigned teams previously in the season.
b. Solve the model you constructed in part (a). Which umpire is assigned to each game? Which umpire is unassigned?

14. **Alternative Optimal Umpire Assignment.** Refer to the umpire assignment problem described in Problem 13. Use the procedure described in Section 14.7 to try to find an alternative optimal solution. **LO 8**

15. **Assigning Dashers to Deliveries.** DoorDash, Inc., headquartered in San Francisco, California, delivers online food orders from restaurants to customers using private vehicles and their drivers, known as Dashers. At any point in time for a given region, DoorDash can offer its drivers available pickups for delivery. The file *doordash* contains data on the amount of time it will take for each of 15 Dashers to reach the restaurant locations of 10 current food orders. **LO 1, 2, 3**

DATA file doordash

a. Formulate a linear programming model to minimize the sum of the times of assignments made such that no Dasher is assigned to more than one order and each order must be assigned to exactly one Dasher.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Solve the model formulated in part (a). What is the minimum total time? Which Dashers are assigned and which are unassigned?

16. **Packaging Materials Vendor Assignment.** United Express Service (UES) uses large quantities of packaging materials at its four distribution hubs. After screening potential suppliers, UES identified six vendors that can provide packaging materials that will satisfy its quality standards. UES asked each of the six vendors to submit bids to satisfy annual demand at each of its four distribution hubs over the next year. The following table lists the bids received (in $1,000s). UES wants to ensure that each of the distribution hubs is serviced by a different vendor. Which bids should UES accept, and which vendors should UES select to supply each distribution hub? **LO 1, 2, 3**

DATA*file* vendors

	Distribution Hub			
Bidder	**1**	**2**	**3**	**4**
Martin Products	190	175	125	230
Schmidt Materials	150	235	155	220
Miller Containers	210	225	135	260
D&J Burns	170	185	190	280
Larbes Furnishings	220	190	140	240
Lawler Depot	270	200	130	260

17. **Scheduling Police Officers.** The Clark County Sheriff's Department schedules police officers for 8-hour shifts. The beginning times for the shifts are 8:00 a.m., noon, 4:00 p.m., 8:00 p.m., midnight, and 4:00 a.m. An officer beginning a shift at one of these times works for the next 8 hours. During normal weekday operations, the number of officers needed varies depending on the time of day. The department staffing guidelines require the following minimum number of officers on duty:

DATA*file* police

Time of Day	**Minimum Number of Officers on Duty**
8:00 a.m.–noon	5
Noon–4:00 p.m.	6
4:00 p.m.–8:00 p.m.	10
8:00 p.m.–midnight	7
Midnight–4:00 a.m.	4
4:00 a.m.–8:00 a.m.	6

Determine the number of police officers who should be scheduled to begin the 8-hour shifts at each of the six times to minimize the total number of officers required. *Hint:* Let x_1 = the number of officers beginning work at 8:00 a.m., x_2 = the number of officers beginning work at noon, and so on. **LO 1, 2**

18. **Fuel Production.** Bay Oil produces two types of fuel (regular and super) by mixing three ingredients. The major distinguishing feature of the two products is the octane level required. Regular fuel must have a minimum octane level of 90, whereas super must have a level of at least 100. The cost per barrel, octane levels, and available amounts (in barrels) for the upcoming two-week period appear in the following table, along with the maximum demand for each end product and the revenue generated per barrel:

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
fuelproduction

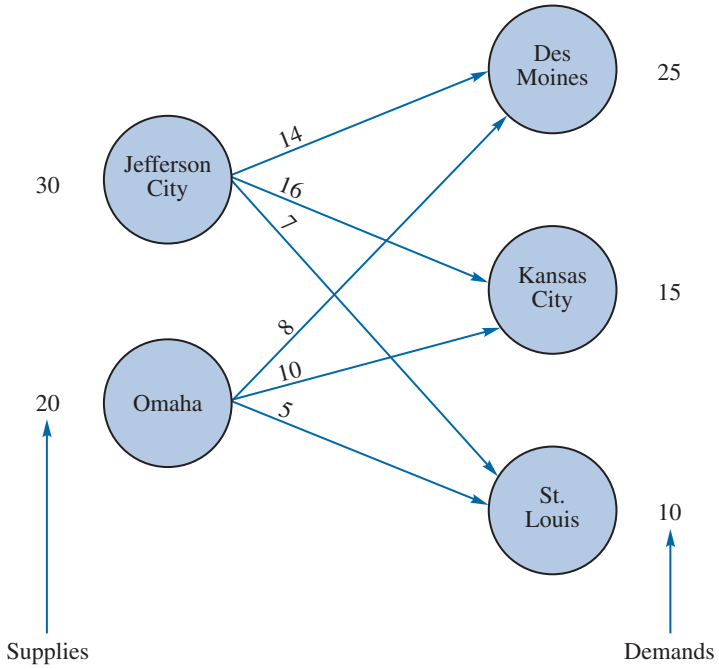
Ingredient	Cost/Barrel ($)	Octane	Available (barrels)
1	16.50	100	110,000
2	14.00	87	350,000
3	17.50	110	300,000

	Revenue/Barrel ($)	Max Demand (barrels)
Regular	18.50	350,000
Super	20.00	500,000

Develop and solve a linear programming model to maximize contribution to profit. What is the optimal contribution to profit? **LO 1, 2, 3**

19. **Distribution Plan to Minimize Cost.** Consider the following network representation of a transportation problem:

DATA *file*
distribution



The supplies, demands, and transportation costs per unit are shown on the network. What is the optimal (cost minimizing) distribution plan? **LO 1, 2, 3**

20. **Alternative Optimal Distribution Plan.** Refer to the transportation problem described in Problem 19. Use the procedure described in Section 14.7 to try to find an alternative optimal solution. **LO 8**

21. **Power Distribution.** Aggie Power Generation supplies electrical power to residential customers for many U.S. cities. Its main power generation plants are located in Los Angeles, Tulsa, and Seattle. The following table shows Aggie Power Generation's major residential markets, the annual demand in each market (in Megawatts or MW), and the cost to supply electricity to each market from each power generation plant (prices are in $/MW). **LO 1, 2, 3**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
aggiepower

Distribution Costs ($/MW)

City	Los Angeles	Tulsa	Seattle	Demand (MW)
Seattle	356.25	593.75	59.38	950.00
Portland	356.25	593.75	178.13	831.25
San Francisco	178.13	475.00	296.88	2,375.00
Boise	356.25	475.00	296.88	593.75
Reno	237.50	475.00	356.25	950.00
Bozeman	415.63	415.63	296.88	593.75
Laramie	356.25	415.63	356.25	1,187.50
Park City	356.25	356.25	475.00	712.50
Flagstaff	178.13	475.00	593.75	1,187.50
Durango	356.25	296.88	593.75	1,543.75

a. If there are no restrictions on the amount of power that can be supplied by any of the power plants, what is the optimal solution to this problem? Which cities should be supplied by which power plants? What is the total annual power distribution cost for this solution?
b. If at most 4,000 MW of power can be supplied by any one of the power plants, what is the optimal solution? What is the annual increase in power distribution cost that results from adding these constraints to the original formulation?

22. **Make Versus Buy for Textiles.** The Calhoun Textile Mill is in the process of deciding on a production schedule. It wishes to know how to weave the various fabrics it will produce during the coming quarter. The sales department has confirmed orders for each of the 15 fabrics produced by Calhoun. These demands are given in the following table. Also given in this table is the variable cost for each fabric. The mill operates continuously during the quarter: 13 weeks, 7 days a week, and 24 hours a day.

There are two types of looms: dobby and regular. Dobby looms can be used to make all fabrics and are the only looms that can weave certain fabrics, such as plaids. The rate of production for each fabric on each type of loom is also given in the table. Note that if the production rate is zero, the fabric cannot be woven on that type of loom. Also, if a fabric can be woven on each type of loom, then the production rates are equal. Calhoun has 90 regular looms and 15 dobby looms. For this problem, assume that the time requirement to change over a loom from one fabric to another is negligible. In addition to producing the fabric using dobby and regular looms, Calhoun has the option to buy some or all of each fabric on the market. The market cost per yard for each fabric is given in the table.

Management would like to know how to allocate the looms to the fabrics and which fabrics to buy on the market so as to minimize the cost of meeting demand. **LO 1, 2, 3**

Fabric	Demand (yd)	Dobby (yd/hr)	Regular (yd/hr)	Mill Cost ($/yd)	Market Cost ($/yd)
1	16,500	4.653	0.00	0.6573	0.80
2	52,000	4.653	0.00	0.5550	0.70
3	45,000	4.653	0.00	0.6550	0.85
4	22,000	4.653	0.00	0.5542	0.70
5	76,500	5.194	5.194	0.6097	0.75

(*Continued*)

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
calhoun

6	110,000	3.809	3.809	0.6153	0.75
7	122,000	4.185	4.185	0.6477	0.80
8	62,000	5.232	5.232	0.4880	0.60
9	7,500	5.232	5.232	0.5029	0.70
10	69,000	5.232	5.232	0.4351	0.60
11	70,000	3.733	3.733	0.6417	0.80
12	82,000	4.185	4.185	0.5675	0.75
13	10,000	4.439	4.439	0.4952	0.65
14	380,000	5.232	5.232	0.3128	0.45
15	62,000	4.185	4.185	0.5029	0.70

23. **Alternative Optimal Plan for Make versus Buy.** Refer to the Calhoun Textile Mill production problem described in Problem 22. Use the procedure described in Section 14.7 to try to find an alternative optimal solution. If you are successful, discuss the differences in the solution you found versus that found in Problem 22. **LO 8**

24. **Fitness Bracelet Distribution.** Orion Fitness produces bracelets with an embedded chip that tracks its wearer's activities. Orion has plants in Denver and Jacksonville. Bracelets produced at either plant may be shipped to either of the firm's two regional warehouses, which are located in Davenport and Cincinnati. These regional warehouses subsequently supply retail outlets in Chicago, Orlando, Houston, and Little Rock. The file *orion* contains data on each plant's supply (number of bracelets), each retail outlet's demand (number of bracelets), and the shipping costs ($ per bracelet) for the shipping channels. **LO 1, 2, 3, 8**

DATA *file*
orion

Shipping Costs ($ per bracelet)

	Warehouse		Supply (number of bracelets)
Plant	Davenport	Cincinnati	
Denver	2.00	3.00	700
Jacksonville	3.00	1.00	400

Shipping Costs ($ per bracelet)

	Retail Outlet			
Warehouse	Chicago	Orlando	Houston	Little Rock
Davenport	3.00	7.00	4.00	7.00
Cincinnati	5.00	5.00	7.00	6.00
Demand (number of bracelets)	200	150	350	300

a. Construct and solve a linear optimization model that defines the supply chain strategy that meets the demand of each retail outlet at minimum total shipping cost.
b. In a separate worksheet, reformulate the problem to determine if there is an alternative optimal solution. Clearly explain your result.

25. **Scheduling Sports Equipment Production.** Brendamore Sports produces footballs and soccer balls and must plan on how many to produce each month for the next six months. The file *brendamore* contains demand forecasts, as well as the production costs and inventory holding costs, for the next six months. Brendamore must meet the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

monthly demand of each product through either a combination of inventory or production during the month. Assume that demand occurs at the end of the month, so that any production during a month can meet that month's demand.

DATA *file* brendamore

Month	Football Demand Forecast	Soccer Ball Demand Forecast
1	15,000	10,000
2	25,000	15,000
3	20,000	10,000
4	5,000	5,000
5	2,500	5,000
6	5,000	7,500

Month	Production Cost ($ per football)	Holding Cost ($ per football)	Production Cost ($ per soccer ball)	Holding Cost ($ per soccer ball)
1	13.80	0.69	10.85	0.54
2	13.90	0.70	10.55	0.53
3	12.95	0.65	10.50	0.53
4	12.60	0.63	10.50	0.53
5	12.55	0.63	10.55	0.53
6	12.70	0.64	10.00	0.50

During each month, there is enough production capacity to produce up to 32,000 total balls (football + soccer balls), and there is enough storage capacity to store up to 20,000 total balls (football + soccer balls) at the end of the month. Brendamore wants to meet these demands on time and it currently has 7,000 footballs and 5,000 soccer balls in inventory. In anticipation of demand beyond the six-month planning horizon, Brendamore would like to have 3,000 footballs and 3,000 soccer balls in inventory at the end of the sixth month.

Brendamore wants to determine the six-month production schedule that minimizes the total production and holding cost. **LO 1, 2, 3**

26. **Developing a Risk-Averse Investment Portfolio.** An investor wishes to invest $10,000 for the coming year and anticipates that the market will be in one of four different states at the end of the year. These states affect her investments in each of three possible stocks and a bond as shown in the following table. Unfortunately, she is uncertain about which market state will occur. Because she is risk-averse, the investor would like to invest in a manner so that the return in the worst-case, no matter what market state occurs, is as good as possible. The following table provides the current price of each possible instrument as well as projected year-end prices of each instrument under each of the four possible states.

DATA *file* marketstates

		Possible Market States			
	Price	State 1	State 2	State 3	State 4
Stock A	$4.94	$4.58	$3.95	$5.67	$5.39
Stock B	$5.88	$5.24	$7.28	$4.82	$6.22
Stock C	$6.48	$8.27	$5.65	$7.66	$5.78
Bond D	$2.68	$2.11	$2.53	$2.80	$2.09

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

These data are in the file *marketstates*. Formulate her investment problem as a linear program and solve it using Excel Solver. How much should she invest in each security? Note: It is possible to purchase fractional shares. **LO 1, 2, 3**

27. **Cash Flow Management.** A financial manager is managing a cash fund. The manager's investment alternatives available are various certificates of deposit, also known as CDs, as listed in the following table:

Investment	Yield	Availability
1-month CD	0.5%	Beginning of each month
3-month CD	1.75%	Beginning of Months 1, 2, 3, 4
6-month CD	2.3%	Beginning of Month 1

DATA *file*
cashmanagement

However, the manager also must ensure that sufficient funds are available to pay company expenditures over the next six months. The following table lists the net expenditures (in thousands of dollars) that the manager is obligated to cover (cash amounts in parenthesis indicate a net inflow of cash rather than outflow).

Month	Net Expenditures ($1000s)
1	4,500
2	(1,100)
3	2,500
4	(2,200)
5	4,300
6	(1,500)

The cash on hand to invest at the start of month 1 is $200,000 and the minimum cash required to be available at the end of month 6 is $100,000. Expenditures are paid at the end of the month. Develop and solve a linear program that will recommend how to invest to maximize the amount of interest income accrued over the next six months while satisfying all financial commitments.

Hint: Investment time starts at the beginning of the month and returns at the end of the month. For example, money invested in a 1-month CD in month 1 will be invested at the beginning of month 1 and returned with interest at the end of month 1. Likewise, money invested in a 3-month CD at the start of month 1 will be returned with interest at the end of month 3. **LO 1, 2, 3**

28. **Organic Food Supply-Chain Optimization.** A produce company supplies organically grown apples to four regional specialty stores. After the apples are collected at the company's orchard, they are transported to any of three preparation centers where they are prepared for retail (by undergoing extensive cleaning and then packaging). After the apples are prepared, they are shipped to the specialty stores. Due to the fragility of the product, the cost of transporting this organic fruit is rather high.

The company has three preparation centers available for use. The file *apple* contains the following: (i) the unit transportation costs (in $/pound) to get the apples from company's orchard to the preparation centers, (ii) the cost ($/pound) to prepare the apples in the preparation centers, (iii) the preparation centers' monthly capacities, (iv) the demand at the specialty stores, and (v) the unit shipping costs of transporting the apples from the preparation centers to the specialty stores. **LO 1, 2, 3, 8**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA file apple

Preparation Center	Transportation Cost ($/pound)	Preparation Cost ($/pound)	Monthly Capacity (pounds)
1	0.45	0.15	300
2	1.00	0.20	500
3	1.62	0.18	800

	Shipping Cost ($ per pound)			
Preparation Center	Organic Orchard	Fresh & Local	Healthy Pantry	Season's Harvest
1	0.80	1.10	0.70	1.40
2	1.20	1.10	0.50	1.40
3	0.20	1.40	1.30	1.70
Monthly Demand (pounds)	300	500	400	200

a. Construct and solve a linear optimization model to determine the number of pounds of apples to prepare at each of the three preparation centers and how much of each specialty store's demand to supply from each preparation center, to minimize the total cost of the operation.
b. In a separate worksheet, reformulate the problem to determine if there is an alternative optimal solution. Clearly explain your result.

29. **Energy Planning.** Windsor Power can store power in a high-capacity battery. The battery has capacity to store 80 kWh (kilowatt hours). During a particular period, Windsor can buy or sell electricity at the market price known as LMP (locational marginal price). The maximum rate that power can be injected or withdrawn from the battery is 20 kWh per period. Windsor has forecasted the following LMPs for the next 10 periods:

DATA file lmp

Period	1	2	3	4	5	6	7	8	9	10
LMP ($/kWh)	5	27	2	25	22	29	24	20	61	66

Develop and solve a linear programming model that will assist Windsor Power in determining how to best buy, store or sell electricity for the next 10 periods based on the LMP. Assume that the battery is half full at the beginning of period 1. That is, at the start of the planning horizon, the battery contains 40 kilowatt hours (kWh) of electricity. It costs $1 to hold 1 kWh in the battery per period. Be sure to define all decision variables. **LO 1, 2, 3**

30. **Ice Cream Recipes.** McKinney's Sweet Shop specializes in homemade candies and ice cream. McKinney's produces its ice cream in house, in batches of 50 pounds. The first stage in ice cream making is the blending of ingredients to obtain a mix which meets pre-specified requirements on the percentages of certain constituents of the mix. The desired composition is as follows:

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Constituent	Required Percentage (%)
1. Fat	16.00
2. Serum Solids	8.00
3. Sugar Solids	16.00
4. Egg Solids	0.35
5. Stabilizer	0.25
6. Emulsifier	0.15
7. Water	59.25

The mix can be composed of ingredients from the following list:

Ingredient	Cost ($/lb)
1. 40% Cream	1.19
2. 23% Cream	0.70
3. Butter	2.32
4. Plastic Cream	2.30
5. Butter Oil	2.87
6. 4% Milk	0.25
7. Skim Condensed Milk	0.35
8. Skim Milk Powder	0.65
9. Liquid Sugar	0.25
10. Sugared Frozen Fresh Egg Yolk	1.75
11. Powdered Egg Yolk	4.45
12. Stabilizer	2.45
13. Emulsifier	1.68
14. Water	0.00

DATA*file*
ice_cream_recipe

The number of pounds of a constituent found in a pound of an ingredient is shown in the following table (the rows correspond to constituents 1 through 7 and the columns correspond to ingredients 1 through 14). Note that a pound of stabilizer contributes only to the stabilizer requirement (one pound), one pound of emulsifier contributes only to the emulsifier requirement (one pound), and that water contributes only to the water requirement (one pound).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.4	0.2	0.8	0.8	0.9	0.1				0.5	0.6			
2	0.1			0.1		0.1	0.3	1						
3									0.7	0.1				
4										0.4	0.4			
5												1		
6													1	
7	0.5	0.8	0.2	0.1	0.1	0.8	0.7		0.3					1

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Jack McKinney Jr. has recently acquired the shop from their father (Papa Jack). Jack's father has in the past used the following mixture: 9.73 pounds of Plastic Cream, 3.03 pounds of Skim Milk Powder, 11.37 pounds of Liquid Sugar, 0.44 pounds of Sugared Frozen Fresh Egg Yolk, 0.12 pounds of Stabilizer, 0.07 pounds of Emulsifier and 25.24 pounds of water. (The scale at McKinney's is only accurate to 100ths of a pound.) **LO 1, 2, 3**

a. Jack feels that perhaps it is possible to produce the ice cream in a more cost-effective manner. Jack would like to find the cheapest mix for producing a batch of ice cream, which meets the requirements specified above. How does the optimal cost compare to the cost of Papa Jack's recipe?
b. Jack is also curious about the cost effect of being a little more flexible in the requirements. Jack wants to know the cheapest mix if the composition meets the following tolerances:

1. Fat	15.00–17.00 %
2. Serum Solids	7.00–9.00 %
3. Sugar Solids	15.50–16.50 %
4. Egg Solids	0.30–0.40 %
5. Stabilizer	0.20–0.30 %
6. Emulsifier	0.10–0.20 %
7. Water	58.00–59.50 %

How much cost benefit is there to being more flexible?

31. **Skateboard Supply Chain.** Sports of All Sorts produces, distributes, and sells high-quality handcrafted skateboards. Its supply chain consists of three factories that produce the skateboards (located in Detroit, Los Angeles, and Austin). The Detroit and Los Angeles facilities can produce 35,000 skateboards per year, but the Austin plant is larger and can produce up to 70,000 skateboards per year. Skateboards must be shipped from the factories to one of four distribution centers, or DCs, located in Iowa, Maryland, Idaho, and Arkansas. Because of differences in labor rates, the processing and handling charge at each DC is different. These costs per skateboard are $1.34 in Iowa, $1.62 in Maryland, $1.30 in Idaho, and $1.38 in Arkansas. Each distribution center can process (repackage, mark for sale, and ship) at most 50,000 skateboards per year. Skateboards are then shipped from the distribution centers to retailers. Sports of All Sorts supplies three major U.S. retailers: Just Sports, Sports 'N Stuff, and The Sports Dude. The following tables display the per unit costs for shipping skateboards between the factories and DCs and for shipping between the DCs and the retailers.

DATA *file* skateboards

	Shipping Costs ($ per skateboard)			
Factory/DCs	**Iowa**	**Maryland**	**Idaho**	**Arkansas**
Detroit	25.00	25.00	35.00	40.00
Los Angeles	35.00	45.00	35.00	42.50
Austin	40.00	40.00	42.50	32.50
Retailers/DC	**Iowa**	**Maryland**	**Idaho**	**Arkansas**
Just Sports	30.00	20.00	35.00	27.50
Sports 'N Stuff	27.50	32.50	40.00	25.00
The Sports Dude	30.00	40.00	32.50	42.50

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This exercise assumes knowledge of concepts discussed in Chapter 9.

Sports of All Sorts needs to forecast the demand at each of its three retail locations for next year, and then plan how to produce and distribute its product from the factories through the DCs to the retailers. The file *skateboards* contains five tabs: Plant Capacities, Transportation Costs, DC Capacities, DC Processing Costs, and Historical Demand. **LO 1, 2, 3**

a. Construct a scatter plot for the historical demand for each retailer. Use the historical demand to forecast the future demand for each retailer. For Just Sports and Sports 'N Stuff, use the average of the historical demands as the forecast for the next year. For The Sports Dude, use simple linear regression to forecast the demand for the next year. Round your forecast to the nearest one thousand units (e.g., if your forecast is 12,303, round to 12,000 for use in part (b).
b. Construct a linear programming model of the supply chain.
c. Solve the linear programming model you constructed in part (b). What is the total cost? Which plants are planned to use all of their capacity? Which distribution centers will use all of their capacity?

Case Problem 1: Investment Strategy

J. D. Williams, Inc. is an investment advisory firm that manages more than $120 million in funds for its numerous clients. The company uses an asset allocation model that recommends the portion of each client's portfolio to be invested in a growth stock fund, an income fund, and a money market fund. To maintain diversity in each client's portfolio, the firm places limits on the percentage of each portfolio that may be invested in each of the three funds. General guidelines indicate that the amount invested in the growth fund must be between 20% and 40% of the total portfolio value. Similar percentages for the other two funds stipulate that between 20% and 50% of the total portfolio value must be in the income fund and that at least 30% of the total portfolio value must be in the money market fund.

In addition, the company attempts to assess the risk tolerance of each client and adjust the portfolio to meet the needs of the individual investor. For example, Williams just contracted with a new client who has $800,000 to invest. Based on an evaluation of the client's risk tolerance, Williams assigned a maximum risk index of 0.05 for the client. The firm's risk indicators show the risk of the growth fund at 0.10, the income fund at 0.07, and the money market fund at 0.01. An overall portfolio risk index is computed as a weighted average of the risk rating for the three funds, where the weights are the fraction of the client's portfolio invested in each of the funds.

Additionally, Williams is currently forecasting annual yields of 18% for the growth fund, 12.5% for the income fund, and 7.5% for the money market fund. Based on the information provided, how should the new client be advised to allocate the $800,000 among the growth, income, and money market funds? Develop a linear programming model that will provide the maximum yield for the portfolio. Use your model to develop a managerial report. **LO 1, 2, 3, 5, 6**

Managerial Report

1. Recommend how much of the $800,000 should be invested in each of the three funds. What is the annual yield you anticipate for the investment recommendation?
2. Assume that the client's risk index could be increased to 0.055. How much would the yield increase, and how would the investment recommendation change?
3. Refer again to the original situation, in which the client's risk index was assessed to be 0.05. How would your investment recommendation change if the annual yield for the growth fund were revised downward to 16% or even to 14%?
4. Assume that the client expressed some concern about having too much money in the growth fund. How would the original recommendation change if the amount invested in the growth fund is not allowed to exceed the amount invested in the income fund?
5. The asset allocation model you developed may be useful in modifying the portfolios for all of the firm's clients whenever the anticipated yields for the three funds are periodically revised. What is your recommendation as to whether use of this model is possible?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Case Problem 2: Solutions Plus

Solutions Plus is an industrial chemicals company that produces specialized cleaning fluids and solvents for a wide variety of applications. Solutions Plus has received an invitation to submit a bid to supply Great North American Railroad with a cleaning fluid for locomotives. Great North American Railroad needs the cleaning fluid at 11 locations (railway stations), and it has provided the following information to Solutions Plus regarding the number of gallons of cleaning fluid required at each location (see the following table).

Solutions Plus can produce the cleaning fluid at its Cincinnati plant for $1.20 per gallon. Even though the Cincinnati location is its only plant, Solutions Plus has negotiated with an industrial chemicals company located in Oakland, California, to produce and ship up to 500,000 gallons of the locomotive cleaning fluid to selected Solutions Plus customer locations.

The Oakland company will charge Solutions Plus $1.65 per gallon to produce the cleaning fluid, but Solutions Plus thinks that the lower shipping costs from Oakland to some customer locations may offset the added cost to produce the product.

DATA*file* solutionsplus

Gallons of Cleaning Fluid Required at Each Location

Location	Gallons Required	Location	Gallons Required
Santa Ana	22,418	Glendale	33,689
El Paso	6,800	Jacksonville	68,486
Pendleton	80,290	Little Rock	148,586
Houston	100,447	Bridgeport	111,475
Kansas City	24,570	Sacramento	112,000
Los Angeles	64,761		

The president of Solutions Plus, Chanda Weaver, contacted several trucking companies to negotiate shipping rates between the two production facilities (Cincinnati and Oakland) and the locations where the railroad locomotives are cleaned. The following table shows the quotes received in terms of dollars per gallon. The "—" entries in the table identify shipping routes that will not be considered because of the large distances involved. These quotes for shipping rates are guaranteed for one year.

	Freight Cost ($ per gallon)	
	Cincinnati	Oakland
Santa Ana	—	0.22
El Paso	0.84	0.74
Pendleton	0.83	0.49
Houston	0.45	—
Kansas City	0.36	—
Los Angeles	—	0.22
Glendale	—	0.22
Jacksonville	0.34	—
Little Rock	0.34	—
Bridgeport	0.34	—
Sacramento	—	0.15

To submit a bid to the railroad company, Solutions Plus must determine the price per gallon it will charge. Solutions Plus usually sells its cleaning fluids for 15% more than

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

its cost to produce and deliver the product. For this big contract, however, Fred Roedel, the director of marketing, suggested that maybe the company should consider a smaller profit margin. In addition, to ensure that Solutions Plus will have adequate capacity to satisfy existing orders as well as accept orders for other new business if it wins the bid, the management team decided to limit the number of gallons of the locomotive cleaning fluid produced in the Cincinnati plant to at most 500,000 gallons. **LO 1, 2, 3**

Managerial Report

You are asked to make recommendations that will help Solutions Plus prepare a bid. Your report should address, but not be limited to, the following issues:

1. If Solutions Plus wins the bid, which production facility (Cincinnati or Oakland) should supply the cleaning fluid to the locations where the railroad locomotives are cleaned? How much should be shipped from each facility to each location?
2. What is the breakeven point for Solutions Plus? That is, what is the lowest bid Solutions Plus can submit without losing money?
3. If Solutions Plus wants to use its standard 15% markup, how much should it bid?
4. Freight costs are significantly affected by the price of fuel. The contract on which Solutions Plus is bidding is for two years. Discuss how fluctuation in freight costs might affect the bid Solutions Plus submits.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 15

Integer Linear Optimization Models

Contents

Analytics in Action: Vattenfall

15.1 Types of Integer Linear Optimization Models

15.2 Eastborne Realty, an Example of Integer Optimization
- The Geometry of Linear All-Integer Optimization

15.3 Solving Integer Optimization Problems with Excel Solver
- A Cautionary Note About Sensitivity Analysis

15.4 Applications Involving Binary Variables
- Capital Budgeting
- Fixed Cost
- Bank Location
- Product Design and Market Share Optimization

15.5 Modeling Flexibility Provided by Binary Variables
- Multiple-Choice and Mutually Exclusive Constraints
- *k* Out of *n* Alternatives Constraint
- Conditional and Corequisite Constraints

15.6 Generating Alternatives in Binary Optimization

Summary 829
Glossary 830
Problems 830

Available in the Cengage eBook:
Appendix: Integer Programming with R

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Formulate an all-integer linear program based on a problem description.

LO 2 Formulate a binary integer linear program based on a problem description.

LO 3 Formulate a mixed-integer linear program based on a problem description.

LO 4 Create a spreadsheet model of all-integer, binary, and mixed-integer linear programs and solve with Excel Solver.

LO 5 Interpret the solution of all-integer, binary, and mixed-integer linear programs.

LO 6 Conduct and interpret a sensitivity analysis for all-integer, binary, and mixed-integer programs.

LO 7 Use binary variables to model multiple choice, mutually exclusive, *k* out of *n* alternative, conditional, and corequisite constraints.

LO 8 Find an alternative optimal solution to a binary or mixed-integer binary linear program when one exists.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

Vattenfall*

European energy company Vattenfall produces electricity in Sweden, Germany, the Netherlands, Denmark, and the United Kingdom. Vattenfall has approximately 20,000 employees and is the second largest offshore wind power developer in the world, investing about 1.5 billion euros annually in wind farms.

Vattenfall uses analytics to optimally design offshore windfarms. Offshore wind farm design involves two major steps; the placement of the turbines that capture the wind and turn it into energy, and the layout of the network of cables that connect the turbines to offshore substations, which collect the energy from the turbines before it is transferred to the main energy grid.

The placement of turbines is an important design issue because turbines can interfere with each other and result in a lower energy production. Called the "wake effect," interference occurs when one turbine changes the airflow and therefore lessons the force of the wind for downstream turbines. Modeling the wake effect involves complex mathematical equations, but Vattenfall uses a linearized version of an optimization model that is a mixed-integer linear program to choose the locations for the turbines. A typical problem might contain 20,000 0-1 variables to represent the choices of turbine placement. Commercial optimization software is used to solve this location problem over a variety of scenarios.

The cable layout problem is solved after the locations for turbines have been decided. The locations of the substations are exogenous to the wind farm planning, that is, they are set by an external organization. The cable layout optimization problem involves deciding how to connect the turbines and the substations to transmit the generated power. A cost minimizing mixed-integer linear program with binary variables indicating connections between turbines and continuous variables indicating the flow of energy is used for this optimization.

The impact of using optimization to design the offshore wind farms is impressive. The models have facilitated increased revenues and decreased costs for Vattenfall, resulting in an annual net gain of approximately 10 million euros per wind farm.

*Based on M. Fischetti et al., "Vattenfall Optimizes Offshore Wind Farm Design," *INFORMS Journal on Applied Analytics*, 50, no. 5 (January–February 2020): 80–94.

In this chapter, we discuss a class of problems that are modeled as linear programs with the additional requirement that one or more variables must be an integer. Such problems are called **integer linear programs**.

The objective of this chapter is to provide an applications-oriented introduction to integer linear programming. We discuss the different types of integer linear programming models, the geometry of all-integer linear programs, and we show how to use Excel Solver to solve integer optimization problems. We discuss four applications of integer linear programming that make use of binary variables: capital budgeting, fixed cost, bank location, and market share optimization problems and we provide additional illustrations of the modeling flexibility provided by binary variables. We also discuss ways to generate useful alternative solutions in integer linear optimization. The online appendix that accompanies this chapter introduces the use of R to solve integer linear programs.

15.1 Types of Integer Linear Optimization Models

The only difference between the problems in this chapter and the problems in Chapter 14 on linear programming is that one or more variables are required to be an integer. If all variables are required to be an integer, we have an **all-integer linear program**. The following is a two-variable, all-integer linear programming model:

$$\begin{aligned} \text{Max} \quad & 2x_1 + 3x_2 \\ \text{s.t.} \quad & \\ & 3x_1 + 3x_2 \leq 12 \\ & \tfrac{2}{3}x_1 + 1x_2 \leq 4 \\ & 1x_1 + 2x_2 \leq 6 \\ & x_1, x_2 \geq 0 \text{ and integer} \end{aligned}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If we drop the phrase *and integer* from the last line of this model, we have the familiar two-variable linear program. The linear program that results from dropping the integer requirements is called the linear programming relaxation, or **LP relaxation**, of the integer linear program.

If some, but not necessarily all, variables are required to be integer, we have a **mixed-integer linear program**. The following is a two-variable, mixed-integer linear program:

$$\begin{aligned} \text{Max} \quad & 3x_1 + 4x_2 \\ \text{s.t.} \quad & \\ & -1x_1 + 2x_2 \leq 8 \\ & 1x_1 + 2x_2 \leq 12 \\ & 2x_1 + 1x_2 \leq 16 \\ & x_1, x_2 \geq 0 \text{ and } x_2 \text{ integer} \end{aligned}$$

We obtain the LP relaxation of this mixed-integer linear program by dropping the requirement that x_2 be integer.

In some applications, the integer variables may take on only the values 0 or 1. Then we have a **binary integer linear program**. As we see later in the chapter, binary variables provide additional modeling capability.

15.2 Eastborne Realty, an Example of Integer Optimization

Eastborne Realty has $2 million available for the purchase of new rental property. After an initial screening, Eastborne reduced the investment alternatives to townhouses and apartment buildings. Each townhouse can be purchased for $282,000, and five are available. Each apartment building can be purchased for $400,000, and the developer will construct as many buildings as Eastborne wants to purchase.

Eastborne's property manager can devote up to 140 hours per month to these new properties; each townhouse is expected to require 4 hours per month, and each apartment building is expected to require 40 hours per month. The annual cash flow, after deducting mortgage payments and operating expenses, is estimated to be $10,000 per townhouse and $15,000 per apartment building. Eastborne's owner would like to determine the number of townhouses and the number of apartment buildings to purchase to maximize annual cash flow.

We begin by defining the decision variables:

$$\begin{aligned} T &= \text{number of townhouses} \\ A &= \text{number of apartment buildings} \end{aligned}$$

The objective function for cash flow (in thousands of dollars) is

$$\text{Max } 10T + 15A$$

Three constraints must be satisfied:

$$\begin{array}{rll} 282T + 400A \leq 2{,}000 & & \text{Funds available (\$1,000s)} \\ 4T + 40A \leq 140 & & \text{Manager's time (hours)} \\ T \leq 5 & & \text{Townhouses available} \end{array}$$

The variables T and A must be nonnegative. In addition, the purchase of a fractional number of townhouses and/or a fractional number of apartment buildings is unacceptable. Thus, T and A must be integers. The model for the Eastborne Realty problem is the following all-integer linear program:

$$\begin{aligned} \text{Max} \quad & 10T + 15A \\ \text{s.t.} \quad & \\ & 282T + 400A \leq 2{,}000 \\ & 4T + 40A \leq 140 \\ & T \leq 5 \\ & T, A \geq 0 \text{ and integer} \end{aligned}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The model for Eastborne Realty is a linear all-integer program. Next we discuss the geometry of this model.

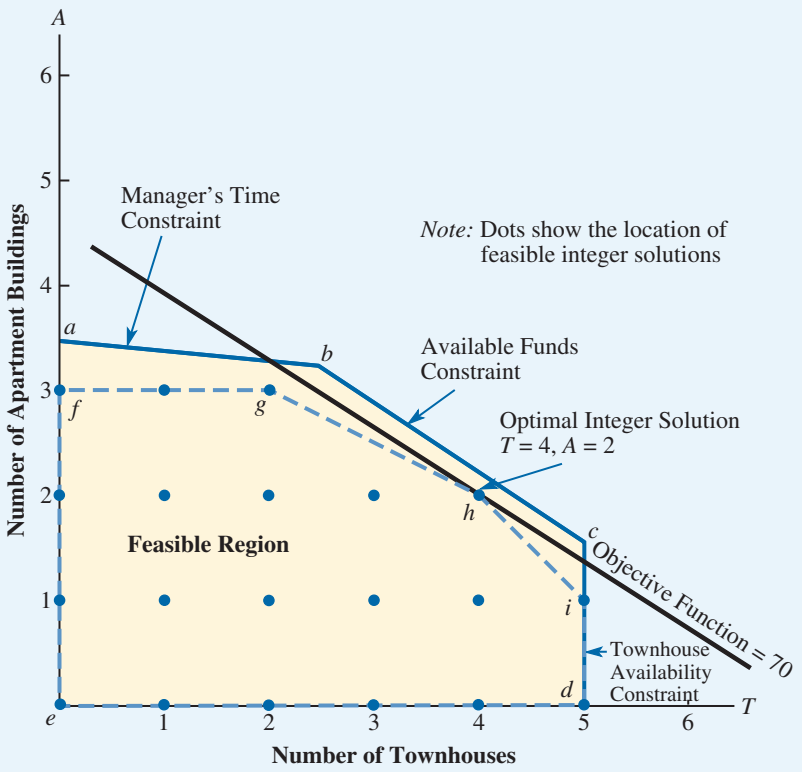
The Geometry of Linear All-Integer Optimization

The geometry of the feasible region for the Eastborne Reality problem is shown in Figure 15.1. The lightly shaded region is the feasible region of the LP relaxation. The optimal linear programming solution is point b, which is $T = 2.479$ townhouses and $A = 3.252$ apartment buildings. The optimal value of the objective function is 73.574, which indicates an annual cash flow of \$73,574. Point b is formed by the intersection of the Manager's Time constraint and the Available Funds constraint. Unfortunately, because Eastborne cannot purchase fractional numbers of townhouses and apartment buildings, further analysis is necessary.

In many cases, a noninteger solution can be rounded to obtain an acceptable integer solution. For instance, a linear programming solution to a production scheduling problem might call for the production of 15,132.4 cases of breakfast cereal. The rounded integer solution of 15,132 cases would probably have minimal impact on the value of the objective function and the feasibility of the solution. Rounding would be a sensible approach. Indeed, whenever rounding has a minimal impact on the objective function and constraints, most managers find it acceptable; a near-optimal solution is satisfactory.

However, rounding may not always be a good strategy. When the decision variables take on small values that have a major impact on the value of the objective function or feasibility, an optimal integer solution is needed. Let us return to the Eastborne Realty problem and examine the impact of rounding. The optimal solution to the LP relaxation for

Figure 15.1 The Geometry of the Eastborne Realty Problem



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Eastborne Realty resulted in $T = 2.479$ townhouses and $A = 3.252$ apartment buildings. Because each townhouse costs \$282,000 and each apartment building costs \$400,000, rounding to an integer solution can be expected to have a substantial economic impact on the problem.

Suppose that we round the solution to the LP relaxation to obtain the integer solution $T = 2$ and $A = 3$, with an objective function value of $10(2) + 15(3) = 65$. The annual cash flow of \$65,000 is substantially less than the annual cash flow of \$73,574 provided by the solution to the LP relaxation. Do other rounding possibilities exist? Exploring other rounding alternatives shows that the integer solution $T = 3$ and $A = 3$ is infeasible because it requires $\$282,000(3) + \$400,000(3) = \$3,738,000$, which is more than the \$2 million that Eastborne has available. The rounded solution of $T = 2$ and $A = 4$ is also infeasible for the same reason. At this point, rounding has led to two townhouses and three apartment buildings with an annual cash flow of \$65,000 as the best feasible integer solution to the problem. Unfortunately, we don't know whether this solution is the best integer solution to the problem.

Rounding to an integer solution is a trial-and-error approach. Each rounded solution must be evaluated for feasibility as well as for its impact on the value of the objective function. Even when a rounded solution is feasible, we have no guarantee that we have found the optimal integer solution. We will see shortly that the rounded solution ($T = 2$ and $A = 3$) is not optimal for Eastborne Realty.

What is the true feasible region for the Eastborne Realty problem? As shown in Figure 15.1, the feasible region is the set of integer points that lie within the feasible region of the LP relaxation. There are 20 such feasible solutions (designated by blue dots in the figure). The region bounded by the dashed lines is known as the **convex hull** of the set of feasible integer solutions. The convex hull of a set of points is the smallest intersection of linear inequalities that contain the set of points. Notice that the convex hull in Figure 15.1 has integer extreme points (points *d*, *e*, *f*, *g*, *h*, and *i*). If we knew the convex hull, we could use linear programming to find the optimal integer corner point. Unfortunately, identifying the convex hull can be very time consuming. This is somewhat counterintuitive because there are only 20 feasible solutions, but solving an integer optimization problem such as that for Eastborne Realty may require solving numerous linear programs to find the optimal integer solution. Therefore, an integer optimization problem can be much more time consuming to solve than solving a linear program of comparable size.

It is true that the optimal solution to the integer program will be an extreme point of the convex hull, so one or more of the extreme points *d*, *e*, *f*, *g*, *h*, and *i* are optimal. The objective function contour shown in Figure 15.1 with an objective function value equal to 70 shows that point *h* is the optimal solution. As a check, let us evaluate each of the corner points of the convex hull in Figure 15.1:

Point	$T =$	$A =$	Annual Cash Flow (\$1,000) =
d	5	0	$10(5) + 15(0) = 50$
e	0	0	$10(0) + 15(0) = 0$
f	0	3	$10(0) + 15(3) = 45$
g	2	3	$10(2) + 15(3) = 65$
h	4	2	$10(4) + 15(2) = 70$
i	5	1	$10(5) + 15(1) = 65$

This confirms that the optimal integer solution occurs at point *h*, where $T = 4$ townhouses and $A = 2$ apartment buildings. The objective function value is an annual cash flow of \$70,000. This solution is substantially better than the best solution found by rounding $T = 2$, $A = 3$ with an annual cash flow of \$65,000. Thus, we see that rounding would not have been the best strategy for Eastborne Realty.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notes + Comments

1. An important observation can be made from the analysis of the Eastborne Realty problem. It has to do with the relationship between the value of the optimal integer solution and the value of the optimal solution to the LP relaxation. For integer linear programs involving maximization, the value of the optimal solution to the LP relaxation provides an upper bound on the value of the optimal integer solution. This observation is valid for the Eastborne Realty problem. The value of the optimal integer solution is $70,000, and the value of the optimal solution to the LP relaxation is $73,574. Thus, we know from the LP relaxation solution that the upper bound for the value of the objective function is $73,574. For integer linear programs involving minimization, the value of the optimal solution to the LP relaxation provides a lower bound on the value of the optimal integer solution.
2. The two popular approaches to solving integer linear optimization problems are branch-and-bound and cutting planes. Both solve a series of LP relaxations to arrive at an optimal integer solution. The *branch-and-bound approach* breaks the feasible region of the LP relaxation into subregions until the subregions have integer solutions or it is determined that the solution cannot be in the subregion. *Cutting plane* approaches try to identify the convex hull by adding a series of new constraints that do not exclude any feasible integer points. Indeed, most software for integer optimization, including Excel Solver, employs a combination of these two approaches.

15.3 Solving Integer Optimization Problems with Excel Solver

The spreadsheet formulation and solution for integer linear programs are similar to that for linear programming problems. Actually the spreadsheet formulation is exactly the same, but some additional information must be provided when setting up the Solver Parameters and Options dialog boxes. Constraints must be added in the Solver Parameters dialog box to identify the integer variables. In addition, the value for Tolerance in the Integer Options dialog box may need to be adjusted to obtain a solution.

Let us demonstrate the Excel solution of an integer linear program by showing how Excel Solver can be used to solve the Eastborne Realty problem. The spreadsheet with the optimal solution is shown in Figure 15.2. We will describe the key elements of the spreadsheet and how to obtain the solution, and then we will interpret the solution.

The parameters and descriptive labels appear in cells A1:G7 of the spreadsheet in Figure 15.2. The cells in the lower portion of the spreadsheet contain the information required by the Excel Solver (decision variables, objective function, constraint left-hand sides, and constraint right-hand sides).

Decision variables	Cells B14:C14 are reserved for the decision variables.
Objective function	The formula =SUMPRODUCT(B7:C7,B14:C14) has been placed into cell B17 to reflect the annual cash flow associated with the solution.
Left-hand sides	The left-hand sides for the three constraints are placed into cells F15:F17. Cell F15 =SUMPRODUCT(B4:C4, $B$14:$C$14) (Copy to cell F16) Cell F17 =B14

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 15.2 Eastborne Realty Spreadsheet Model

	A	B	C	D	E	F	G
1	**Eastborne Realty Problem**						
2							
3		**Townhouse**	**Apt. Bldg.**				
4	**Price (000)**	282	400			**Funds Avl. ($000)**	2000
5	**Mgr. Time**	4	40			**Mgr. Time Avl. (Hours)**	140
6						**Townhouses Avl.**	5
7	**Ann. Cash Flow ($000)**	10	15				
8							
9							
10	**Model**						
11							
12		Number of					
13		Townhouses	Apt. Bldgs.				
14	**Purchase Plan**	4	2			Total Used	Total Available
15					Funds (000)	=SUMPRODUCT(B4:C4,$B$14:$C$14)	=G4
16					Time (Hours)	=SUMPRODUCT(B5:C5,$B$14:$C$14)	=G5
17	**Max Cash Flow (000)**	=SUMPRODUCT(B7:C7,B14:C14)			Townhouses	=B14	=G6
18							

MODEL *file*
eastborne

	A	B	C	D	E	F	G
1	**Eastborne Realty Problem**						
2							
3		**Townhouse**	**Apt. Bldg.**				
4	**Price (000)**	$282	$400			**Funds Avl. ($000)**	$2,000
5	**Mgr. Time**	4	40			**Mgr. Time Avl. (Hours)**	140
6						**Townhouses Avl.**	5
7	**Ann. Cash Flow ($000)**	$10	$15				
8							
9							
10	**Model**						
11							
12		Number of					
13		Townhouses	Apt. Bldgs.				
14	**Purchase Plan**	4	2			Total Used	Total Available
15					Funds (000)	$1,928	$2,000
16					Time (Hours)	96	140
17	**Max Cash Flow (000)**	$70			Townhouses	4	5
18							

Right-hand sides The right-hand sides for the three constraints are placed into cells G15:G17.
Cell G15 =G4 (Copy to cells G16:G17)

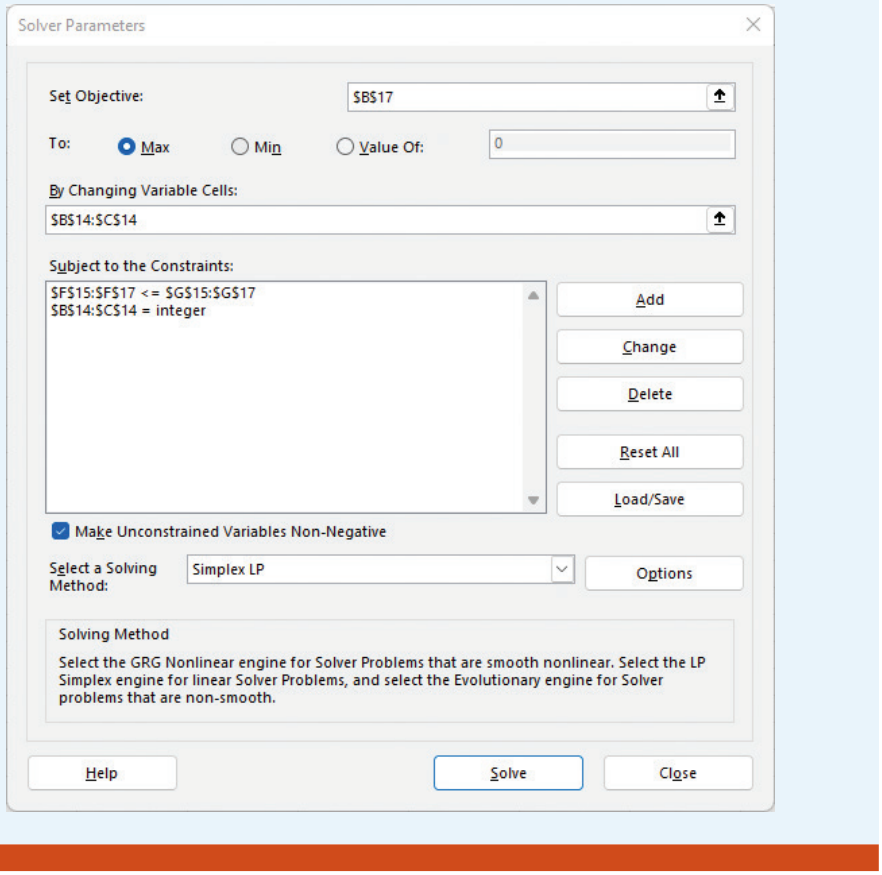
To solve the Eastborne Realty problem, we follow these steps:

Step 1. Click the **Data** tab in the Ribbon
Step 2. In the **Analyze** group, click **Solver**
Step 3. When the **Solver Parameters** dialog box appears (Figure 15.3):
Enter *B17* in the **Set Objective:** box
Select **Max** for the **To:** option
Enter *B14:C14* in the **By Changing Variable Cells:** box
Step 4. Click the **Add** button
When the **Add Constraint** dialog box appears:
Enter *B14:C14* in the **Cell Reference:** box
Select **int** from the drop-down menu
When **int** is selected, the term "integer" automatically appears in the **Constraint:** box. This constraint tells Solver that the decision variables in cells B14 and C14 must be integers.

*Binary variables are identified with the **bin** designation in the **Add Constraint** dialog box.*

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 15.3 Solver Parameters Dialog Box for Eastborne Realty



Step 5. Click the **Add** button
When the **Add Constraint** dialog box appears:
Enter *F15:F17* in the **Cell Reference:** box
Select ≤ from the drop-down menu
Enter *G15:G17* in the **Constraint:** area
Click **OK**

Step 6. Select the **Make Unconstrained Variables Non-Negative** option
Select **Simplex LP** from the **Select a Solving Method:** drop-down menu

Step 7. Click the **Options** button
Select the **All Methods** tab, and set the **Integer Optimality (%):** to *0*, as shown in Figure 15.4. This ensures that we find the optimal integer solution
Click **OK** to close the **Options** dialog box

Step 8. When the **Solver Parameters** dialog box reappears, click **Solve**

Step 9. When the **Solver Results** dialog box appears, select **Answer** in the **Reports** area and click **OK**

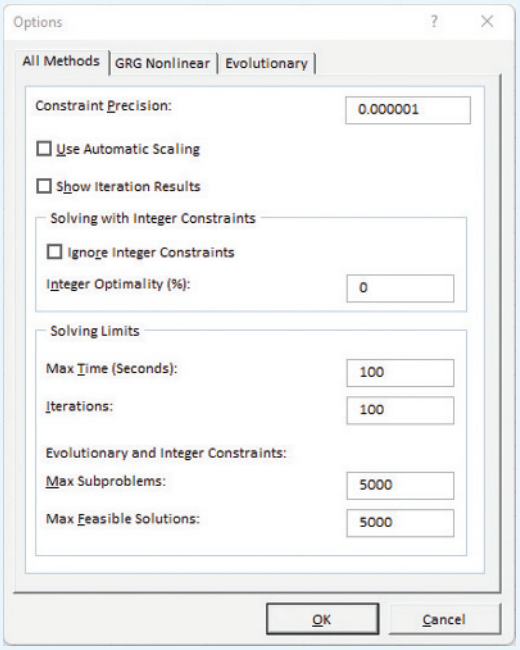
MODEL *file*
eastbornemodel

The completed linear integer optimization model for the Eastborne Realty problem is contained in the file *eastbornemodel.*

Figure 15.5 shows the Eastborne Realty Answer Report. The structure of the Answer Report from Excel Solver for integer programs is the same as that described

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 15.4 Solver Options Dialog Box



in Chapter 14 for linear programs. The first section gives information regarding the objective function. It shows that the objective function is located in cell B17 and that the optimal value (Final Value) of the objective function is $70,000. The Variable Cells section gives the location, name, and original and optimal values (Final Value) of the decision variables, as well as an indication that the decision variables have been designated as integers. For the Eastborne problem, in Figure 15.5, we see that the optimal solution is to purchase four townhouses and two apartment buildings. Finally, the Constraints section gives us detail on the status of each constraint at optimality. We see that none of the three constraints is binding, and from the slack column, we see that we have $72,000 unused from budget and 44 unused hours and that we are under the limit of 5 townhouses by 1.

As this example illustrates, and as we have seen in Figure 15.1, unlike in a linear program, the solution to an integer program can be such that none of the constraints are binding at the optimal point.

A Cautionary Note About Sensitivity Analysis

The classical sensitivity analysis discussed in Chapter 14 for linear programs is not available for integer programs. Because of the discrete nature of integer optimization, it is not possible to easily calculate objective function coefficient ranges, shadow prices, and right-hand-side ranges. However, this does not mean that the sensitivity analysis is not important for integer programs. Sensitivity analysis is often more crucial for integer linear programming problems than for linear programming problems. A small change in one of the coefficients in the constraints can cause a relatively large change in the value of the optimal solution. To understand why, consider the following integer programming model of

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 15.5 Excel Solver Answer Report for the Eastborne Realty Problem

	A	B	C	D	E	F	G
13							
14	Objective Cell (Max)						
15		Cell	Name	Original Value	Final Value		
16		$B$17	Max Cash Flow (000) Townhouses	$0	$70		
17							
18							
19	Variable Cells						
20		Cell	Name	Original Value	Final Value	Integer	
21		$B$14	Purchase Plan Townhouses	0	4	Integer	
22		$C$14	Purchase Plan Apt. Bldgs.	0	2	Integer	
23							
24							
25	Constraints						
26		Cell	Name	Cell Value	Formula	Status	Slack
27		$F$15	Funds (000) Total Used	$1,928	$F$15<=$G$15	Not Binding	72
28		$F$16	Time (Hours) Total Used	96	$F$16<=$G$16	Not Binding	44
29		$F$17	Townhouses Total Used	4	$F$17<=$G$17	Not Binding	1
30		$B$14:$C$14=Integer					
31							

a simple capital budgeting problem involving four projects and a budgetary constraint for a single time period:

$$\begin{aligned} \text{Max} \quad & 40x_1 + 60x_2 + 70x_3 + 160x_4 \\ \text{s.t.} \quad & \\ & 16x_1 + 35x_2 + 45x_3 + 85x_4 \leq 100 \\ & x_1, x_2, x_3, x_4 = 0, 1 \end{aligned}$$

Sensitivity reports are not available for integer optimization problems. To determine the sensitivity of the solution to changes in model inputs, you must change the data and re-solve the problem.

The optimal solution to this problem is $x_1 = 1$, $x_2 = 1$, $x_3 = 1$, and $x_4 = 0$, with an objective function value of \$170. However, if the available budget is increased by \$1 (from \$100 to \$101), the optimal solution changes to $x_1 = 1$, $x_2 = 0$, $x_3 = 0$, and $x_4 = 1$, with an objective function value of \$200. In other words, one additional dollar in the budget would lead to a \$30 increase in the return. Surely management, when faced with such a situation, would increase the budget by \$1. Because of the extreme sensitivity of the value of the optimal solution to the constraint coefficients, practitioners usually recommend re-solving the integer linear program several times with variations in the coefficients before attempting to choose the best solution for implementation.

Notes + Comments

The time required to obtain an optimal solution can be highly variable for integer linear programs. If an optimal solution cannot be found within a reasonable amount of time, the **Integer Optimality (%)** can be reset to 5% or some higher value so that the search procedure may stop when a near-optimal solution (within the tolerance of being optimal) has been found. This can shorten the solution time because, if the **Integer Optimality (%)** is set to 5%, Solver can stop when it knows it is within 5% of optimal rather than having to complete the search. In general, unless you are experiencing excessive run times, we recommend you set the **Integer Optimality (%)** to 0.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

15.4 Applications Involving Binary Variables

Much of the modeling flexibility provided by integer linear programming is due to the use of binary variables. In many applications, binary variables provide selections or choices with the value of the variable equal to one if a corresponding activity is undertaken and equal to zero if the corresponding activity is not undertaken. The capital budgeting, fixed cost, bank location, and product design and market share optimization applications presented in this section make use of binary variables.

Capital Budgeting

The estimated net present value is the net cash flow discounted back to the beginning of year 1.

The Ice-Cold Refrigerator Company is considering investing in several projects that have varying capital requirements over the next four years. Faced with limited capital each year, management would like to select the most profitable projects that it can afford. The estimated net present value for each project, the capital requirements, and the available capital over the four-year period are shown in Table 15.1.

Let us define four binary decision variables:

P = 1 if the plant expansion project is accepted; 0 if rejected
W = 1 if the warehouse expansion project is accepted; 0 if rejected
M = 1 if the new machinery project is accepted; 0 if rejected
R = 1 if the new product research project is accepted; 0 if rejected

In a **capital budgeting problem**, the company's objective function is to maximize the net present value of the capital budgeting projects. This problem has four constraints: one for the funds available in each of the next four years.

A binary integer linear programming model with dollars in thousands is as follows:

$$\begin{aligned}
\text{Max} \quad & 90P + 40W + 10M + 37R \\
\text{s.t.} \quad & \\
& 15P + 10W + 10M + 15R \le 40 \quad \text{(Year 1 capital available)} \\
& 20P + 15W \qquad + 10R \le 50 \quad \text{(Year 2 capital available)} \\
& 20P + 20W \qquad + 10R \le 40 \quad \text{(Year 3 capital available)} \\
& 15P + 5W + 4M + 10R \le 35 \quad \text{(Year 4 capital available)} \\
& P, W, M, R = 0, 1
\end{aligned}$$

The Ice-Cold spreadsheet model and Solver dialog box are shown in Figure 15.6. The SUMPRODUCT function is used to calculate the amount of capital used in each year as well as the net present value.

The Excel Solver Answer Report is shown in Figure 15.7. The optimal solution is $P = 1, W = 1, M = 1, R = 0$, with a total estimated net present value of $140,000.

Table 15.1 Project Net Present Value, Capital Requirements, and Available Capital for the Ice-Cold Refrigerator Company

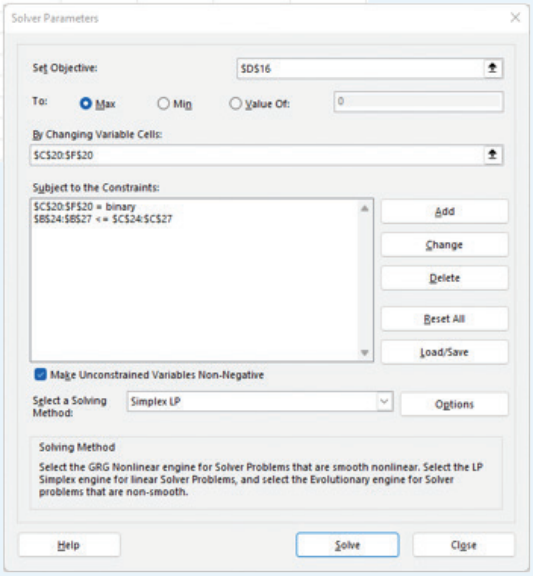
	Project				
	Plant Expansion ($)	Warehouse Expansion ($)	New Machinery ($)	New Product Research ($)	Total Capital Available ($)
Present Value	90,000	40,000	10,000	37,000	
Year 1 Cap Rqmt	15,000	10,000	10,000	15,000	40,000
Year 2 Cap Rqmt	20,000	15,000		10,000	50,000
Year 3 Cap Rqmt	20,000	20,000		10,000	40,000
Year 4 Cap Rqmt	15,000	5,000	4,000	10,000	35,000

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 15.6 Ice-Cold Spreadsheet Model and Solver Dialog Box

	A	B	C	D	E	F	G	H
1	Ice-Cold Refrigerator							
2	Parameters							
3				Financial Data ($1000s)				
4			Plant	Warehouse	New	New Prod.		
5			Expansion	Expansion	Machinery	Research	Capital	
6		Net Present Value	$90	$40	$10	$37	Available	
7		Year 1 Capital	$15	$10	$10	$15	$40	
8		Year 2 Capital	$20	$15		$10	$50	
9		Year 3 Capital	$20	$20		$10	$40	
10		Year 4 Capital	$15	$5	$4	$10	$35	
11								
12								
13	Model							
14								
15								
16		Max Net Present Value ($1000s)		$140.00				
17								
18			Plant	Warehouse	New	New Prod.		
19			Expansion	Expansion	Machinery	Research		
20		Investment Plan	1	1	1	0		
21								
22		Amount ($1000s)						
23		Spent	Available					
24	Year 1	$35	$40					
25	Year 2	$35	$50					
26	Year 3	$40	$40					
27	Year 4	$24	$35					
28								

MODEL file
icecold



Thus, the company should fund the plant expansion, warehouse expansion, and new machinery projects. The new product research project should be put on hold unless additional capital funds become available. The values of the slack variables (Figure 15.7) show that the company will have $5,000 remaining in year 1, $15,000 remaining in year 2, and $11,000 remaining in year 4. Checking the capital requirements for the new product research project, we see that enough funds are available for this project in years 2 and 4. However, the company would have to find additional capital funds of $10,000 in year 1 and $10,000 in year 3 to fund the new product research project.

Fixed Cost

In many applications, the cost of production has two components: a fixed setup cost and a variable cost directly related to the production quantity. The use of binary

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 15.7 Answer Report for Ice-Cold Refrigerator

Objective Cell (Max)

Cell	Name	Original Value	Final Value
$D$16	Max Net Present Value ($1000s) Expansion	$0.00	$140.00

Variable Cells

Cell	Name	Original Value	Final Value	Integer
$C$20	Investment Plan Expansion	0	1	Binary
$D$20	Investment Plan Expansion	0	1	Binary
$E$20	Investment Plan Machinery	0	1	Binary
$F$20	Investment Plan Research	0	0	Binary

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$B$24	Year 1 Spent	$35	$B$24<=$C$24	Not Binding	5
$B$25	Year 2 Spent	$35	$B$25<=$C$25	Not Binding	15
$B$26	Year 3 Spent	$40	$B$26<=$C$26	Binding	0
$B$27	Year 4 Spent	$24	$B$27<=$C$27	Not Binding	11

$C$20:$F$20=Binary

variables makes including the setup cost possible in a model for a production application.

As an example of a **fixed-cost problem**, consider the production problem faced by RMC Inc. Three raw materials are used to produce three products: a fuel additive, a solvent base, and a carpet cleaning fluid. The following decision variables are used:

F = tons of fuel additive produced
S = tons of solvent base produced
C = tons of carpet cleaning fluid produced

The profit contributions are $40 per ton for the fuel additive, $30 per ton for the solvent base, and $50 per ton for the carpet cleaning fluid. Each ton of fuel additive is a blend of 0.4 ton of material 1 and 0.6 ton of material 3. Each ton of solvent base requires 0.5 ton of material 1, 0.2 ton of material 2, and 0.3 ton of material 3. Each ton of carpet cleaning fluid is a blend of 0.6 ton of material 1, 0.1 ton of material 2, and 0.3 ton of material 3. RMC has 20 tons of material 1, 5 tons of material 2, and 21 tons of material 3, and management is interested in determining the optimal production quantities for the upcoming planning period.

A linear programming model of the RMC problem is as follows:

$$\begin{aligned}
&\text{Max } 40F + 30S + 50C \\
&\text{s.t.} \\
&\quad 0.4F + 0.5S + 0.6C \le 20 \quad \text{Material 1} \\
&\quad 0.2S + 0.1C \le 5 \quad \text{Material 2} \\
&\quad 0.6F + 0.3S + 0.3C \le 21 \quad \text{Material 3} \\
&\quad F,S,C \ge 0
\end{aligned}$$

Using Excel Solver, we obtain an optimal solution consisting of 27.5 tons of fuel additive, 0 tons of solvent base, and 15 tons of carpet cleaning fluid, with a value of $1,850.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This linear programming formulation of the RMC problem does not include a fixed cost for production setup of the products. Suppose that the following data are available concerning the setup cost and the maximum production quantity for each of the three products:

Product	Setup Cost ($)	Maximum Production (tons)
Fuel additive	200	50
Solvent base	50	25
Carpet cleaning fluid	400	40

The modeling flexibility provided by binary variables can now be used to incorporate the fixed setup costs into the production model. The binary variables are defined as follows:

$SF = 1$ if the fuel additive is produced; 0 if not
$SS = 1$ if the solvent base is produced; 0 if not
$SC = 1$ if the carpet cleaning fluid is produced; 0 if not

Using these setup variables, the total setup cost is

$$200SF + 50SS + 400SC$$

We can now rewrite the objective function to include the setup cost. Thus, the net profit objective function becomes

$$\text{Max } 40F + 30S + 50C - 200SF - 50SS - 400SC$$

Next, we must write production capacity constraints so that, if a setup variable equals 0, production of the corresponding product is not permitted, and if a setup variable equals 1, production is permitted up to the maximum quantity. For the fuel additive, we do so by adding the following constraint:

$$F \leq 50SF$$

Note that, with this constraint present, production of the fuel additive is not permitted when $SF = 0$. When $SF = 1$, production of up to 50 tons of fuel additive is permitted. We can think of the setup variable as a switch. When it is off ($SF = 0$), production is not permitted; when it is on ($SF = 1$), production is permitted.

Similar production capacity constraints, using binary variables, are added for the solvent base and carpet cleaning products:

$$S \leq 25SS$$
$$C \leq 40SC$$

In summary, we have the following fixed-cost model for the RMC problem with setups:

$$\text{Max } 40F + 30S + 50C - 200SF - 50SS - 400SC$$
s.t.

$$\begin{aligned}
0.4F + 0.5S + 0.6C &\leq 20 && \text{Material 1}\\
0.2S + 0.1C &\leq 5 && \text{Material 2}\\
0.6F + 0.3S + 0.3C &\leq 21 && \text{Material 3}\\
F &\leq 50SF && \text{Maximum Fuel Additive}\\
S &\leq 25SS && \text{Maximum Solvent Base}\\
C &\leq 40SC && \text{Maximum Carpet Cleaning}
\end{aligned}$$

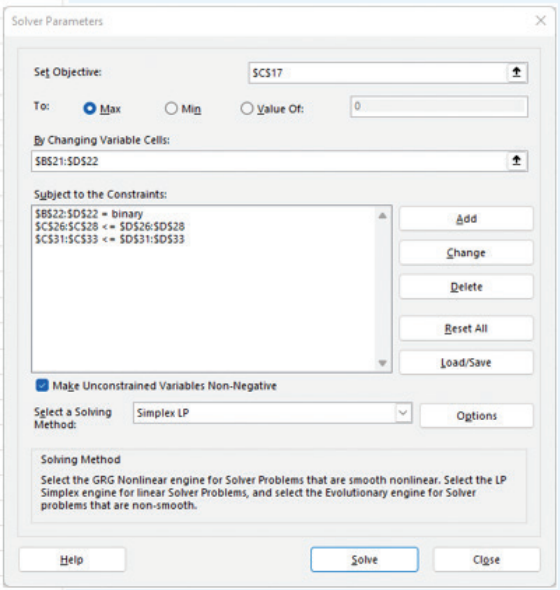
$$F, S, C \geq 0;\ SF, SS, SC = 0 \text{ or } 1$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 15.8 RMC with Setups Spreadsheet Model and Solver Dialog Box

	A	B	C	D	E
1	**RMC**				
2	**Parameters**				
3			Material Requirements (tons)		
4		Fuel	Solvent	Cleaning	Tons
5	Materials	Additive	Base	Fluid	Available
6	Material 1	0.4	0.5	0.6	20
7	Material 2		0.2	0.1	5
8	Material 3	0.6	0.3	0.3	21
9	Profit per ton	40	30	50	
10	Setup Cost	200	50	400	
11	Capacity (Tons)	50	25	40	
12					
13					
14	**Model**				
15					
16					
17		Max Net Profit	$1,350.00		
18					
19					
20		Fuel	Solvent	Cleaning	
21	Tons Produced	25.0	20.0	0.0	
22	Setup	1	1	0	
23					
24					
25			Used	Available	
26		Material 1	20	20	
27		Material 2	4	5	
28		Material 3	21	21	
29					
30			Tons Produced	Max Tons	
31		Max F	25	50	
32		Max S	20	25	
33		Max C	0.0	0	

MODEL *file*
rmcsetup



A spreadsheet model and Solver dialog box for the RMC problem are shown in Figure 15.8. The SUMPRODUCT function is used to calculate the material used, and cells D31, D32, and D33 contain the capacity multiplied by the appropriate binary variable (=B11*B22 in cell D31, =C11*C22 in cell D32, and =D11*D22 in cell D33).

The Excel Answer Report is shown in Figure 15.9. The optimal solution requires 25 tons of fuel additive and 20 tons of solvent base. The value of the objective function after deducting the setup cost is $1,350. The setup cost for the fuel additive and the solvent base is $200 + $50 = $250. The optimal solution includes $SC = 0$, which indicates that the more expensive $400 setup cost for the carpet cleaning fluid should be avoided. Thus, the carpet cleaning fluid is not produced.

The key to developing a fixed-cost model is the introduction of a binary variable for each fixed cost and the specification of an upper bound for the corresponding production variable. For a production quantity x, a constraint of the form $x \leq My$ can then be used to allow production when the setup variable $y = 1$ and not to allow production when the setup variable $y = 0$. The value of the maximum production quantity M should be large enough to allow for all reasonable levels of production, but choosing excessively large values of M will slow the solution procedure.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 15.9 Answer Report for RMC Production Problem

Objective Cell (Max)

Cell	Name	Original Value	Final Value
$C$17	Max Net Profit Base	$0.00	$1,350.00

Variable Cells

Cell	Name	Original Value	Final Value	Integer
$B$21	Tons Produced Fuel	0.0	25.0	Contin
$C$21	Tons Produced Solvent	0.0	20.0	Contin
$D$21	Tons Produced Cleaning	0.0	0.0	Contin
$B$22	Setup Fuel	0	1	Binary
$C$22	Setup Solvent	0	1	Binary
$D$22	Setup Cleaning	0	0	Binary

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$C$26	Material 1 Used	20	$C$26<=$D$26	Binding	0
$C$27	Material 2 Used	4	$C$27<=$D$27	Not Binding	1
$C$28	Material 3 Used	21	$C$28<=$D$28	Binding	0
$C$31	Max F Tons Produced	25	$C$31<=$D$31	Not Binding	25
$C$32	Max S Tons Produced	20	$C$32<=$D$32	Not Binding	5
$C$33	Max C Tons Produced	0.0	$C$33<=$D$33	Binding	0

$B$22:$D$22=Binary

Bank Location

The long-range planning department for the Ohio Trust Company is considering expanding its operation into a 20-county region in northeastern Ohio (Figure 15.10). Currently, Ohio Trust does not have a principal place of business in any of the 20 counties. According to the banking laws in Ohio, if a bank establishes a principal place of business (PPB) in any county, branch banks can be established in that county and in any of the adjacent counties. However, to establish a new principal place of business, Ohio Trust must either obtain approval for a new bank from the state's superintendent of banks or purchase an existing bank.

Table 15.2 lists the 20 counties in the region and adjacent counties. For example, Ashtabula County is adjacent to Lake, Geauga, and Trumbull counties; Lake County is adjacent to Ashtabula, Cuyahoga, and Geauga counties; and so on.

As an initial step in its planning, Ohio Trust would like to determine the minimum number of PPBs necessary to do business throughout the 20-county region. A binary integer programming model can be used to solve this **location problem** for Ohio Trust. We define the variables as

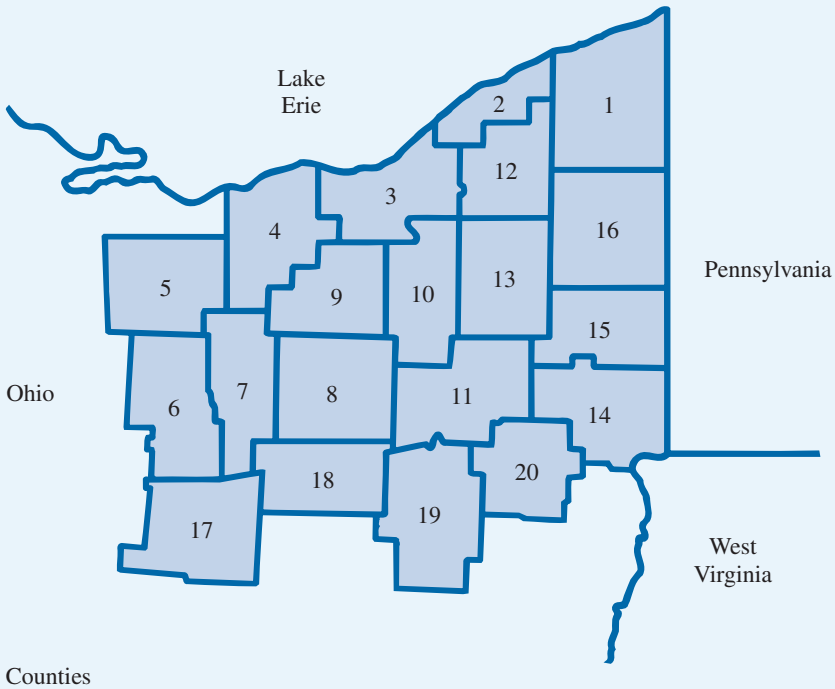
$$x_i = 1 \text{ if a PPB is established in county } i\text{; 0 otherwise}$$

To minimize the number of PPBs needed, we write the objective function as

$$\text{Min } x_1 + x_2 + \cdots + x_{20}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 15.10 Ohio Trust County Map



Counties

1. Ashtabula	6. Richland	11. Stark	16. Trumbull
2. Lake	7. Ashland	12. Geauga	17. Knox
3. Cuyahoga	8. Wayne	13. Portage	18. Holmes
4. Lorain	9. Medina	14. Columbiana	19. Tuscarawas
5. Huron	10. Summit	15. Mahoning	20. Carroll

The bank may locate branches in a county if the county contains a PPB or is adjacent to another county with a PPB. Thus, the binary linear program will need one constraint for each county. For example, the constraint for Ashtabula County is

$$x_1 + x_2 + x_{12} + x_{16} \geq 1 \quad \text{Ashtabula}$$

MODEL *file*
ohiotrust

Note that satisfaction of this constraint ensures that a PPB will be placed in Ashtabula County *or* in one or more of the adjacent counties. This constraint thus guarantees that Ohio Trust will be able to place branch banks in Ashtabula County.

The statement of the bank location problem is as follows:

$$\begin{aligned}
\text{Min} \quad & x_1 + x_2 + \quad \ldots \quad + x_{20} \\
\text{s.t.} \quad & \\
& x_1 + x_2 \quad + x_{12} + x_{16} \quad \geq 1 \quad \text{Ashtabula} \\
& x_1 + x_2 + x_3 + x_{12} \quad \geq 1 \quad \text{Lake} \\
& \vdots \qquad\qquad\qquad\qquad \vdots \\
& \qquad\qquad x_{11} + x_{14} + x_{19} + x_{20} \quad \geq 1 \quad \text{Carroll} \\
& x_i = 0, 1 \quad i = 1, 2, \ldots, 20
\end{aligned}$$

Note that we have only listed the constraints for counties 1, 2, and 20 (Ashtabula, Lake, and Carroll). Constraints for counties 3-19 are constructed in a similar manner.

We use Excel Solver to solve this 20-variable, 20-constraint problem formulation. In Figure 15.11, we show the optimal solution. The optimal solution calls for principal places of business in Ashland, Stark, and Geauga counties. With PPBs in these three counties, Ohio Trust can place branch banks in all 20 counties. Clearly the integer programming model could be enlarged to allow for expansion into a larger area or throughout the entire state.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 15.2 Counties in the Ohio Trust Expansion Region

Counties Under Consideration	Adjacent Counties (by Number)
1. Ashtabula	2, 12, 16
2. Lake	1, 3, 12
3. Cuyahoga	2, 4, 9, 10, 12, 13
4. Lorain	3, 5, 7, 9
5. Huron	4, 6, 7
6. Richland	5, 7, 17
7. Ashland	4, 5, 6, 8, 9, 17, 18
8. Wayne	7, 9, 10, 11, 18
9. Medina	3, 4, 7, 8, 10
10. Summit	3, 8, 9, 11, 12, 13
11. Stark	8, 10, 13, 14, 15, 18, 19, 20
12. Geauga	1, 2, 3, 10, 13, 16
13. Portage	3, 10, 11, 12, 15, 16
14. Columbiana	11, 15, 20
15. Mahoning	11, 13, 14, 16
16. Trumbull	1, 12, 13, 15
17. Knox	6, 7, 18
18. Holmes	7, 8, 11, 17, 19
19. Tuscarawas	11, 18, 20
20. Carroll	11, 14, 19

Product Design and Market Share Optimization

Conjoint analysis is a market research technique that can be used to learn how prospective buyers of a product value the product's attributes. In this section, we will show how the results of conjoint analysis can be used in an integer programming model of a **product design and market share optimization problem**. We illustrate the approach by considering a problem facing Salem Foods, a major producer of frozen foods.

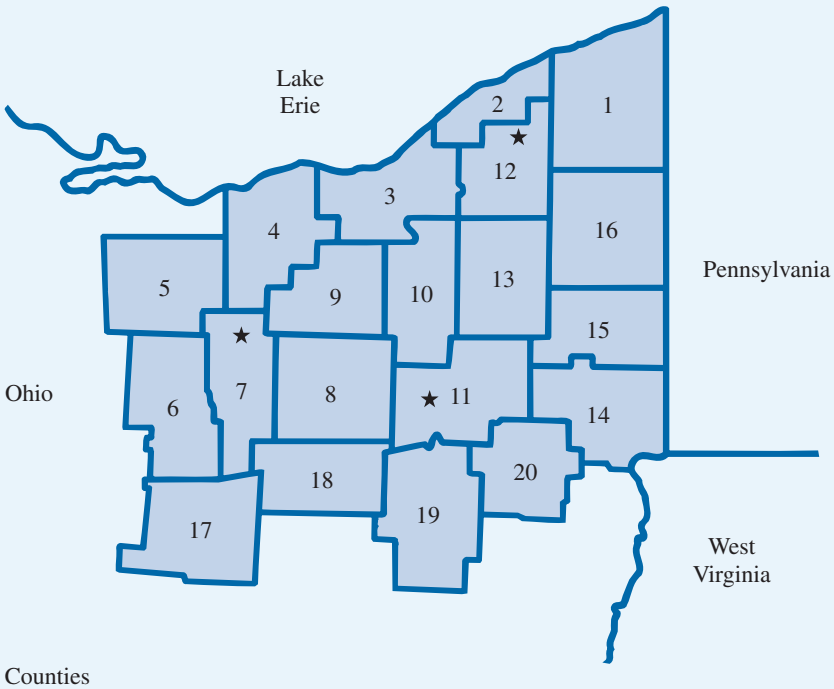
Salem Foods is planning to enter the frozen pizza market. Currently, two existing brands, Antonio's and King's, have the major share of the market. In trying to develop a sausage pizza that will capture a substantial share of the market, Salem determined that the four most important attributes when consumers purchase a frozen sausage pizza are crust, cheese, sauce, and sausage flavor. The crust attribute has two levels (thin and thick); the cheese attribute has two levels (mozzarella and blend); the sauce attribute has two levels (smooth and chunky); and the sausage flavor attribute has three levels (mild, medium, and hot).

In a typical conjoint analysis, a sample of consumers is asked to express their preference for a product with chosen levels for the attributes. Then regression analysis is used to determine the part-worth for each of the attribute levels. In essence, the **part-worth** is the utility value that a consumer attaches to each level of each attribute. Provided part-worths from regression analysis, we will show how they can be used to determine the overall value a consumer attaches to a particular product.

Table 15.3 shows the part-worths for each level of each attribute provided by a sample of eight potential Salem customers who are currently buying either King's or Antonio's pizza. For consumer 1, the part-worths for the crust attribute are 11 for thin crust and 2 for thick crust, indicating a preference for thin crust. For the cheese attribute, the part-worths

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 15.11 Optimal Solution to the Ohio Trust Location Problem



Counties

1. Ashtabula
2. Lake
3. Cuyahoga
4. Lorain
5. Huron
6. Richland
7. Ashland
8. Wayne
9. Medina
10. Summit
11. Stark
12. Geauga
13. Portage
14. Columbiana
15. Mahoning
16. Trumbull
17. Knox
18. Holmes
19. Tuscarawas
20. Carroll

★ *A principal place of business should be located in these counties.*

are 6 for the mozzarella cheese and 7 for the cheese blend; thus, consumer 1 has a slight preference for the cheese blend. From the other part-worths, we see that consumer 1 shows a strong preference for the chunky sauce over the smooth sauce (17 to 3) and has a slight preference for the medium-flavored sausage. Note that consumer 2 shows a preference for the thin crust, the cheese blend, the chunky sauce, and mild-flavored sausage. The part-worths for the other consumers are interpreted similarly.

Table 15.3 Part-Worths for the Salem Foods Problem

	Crust		Cheese		Sauce		Sausage Flavor		
Consumer	Thin	Thick	Mozzarella	Blend	Smooth	Chunky	Mild	Medium	Hot
1	11	2	6	7	3	17	26	27	8
2	11	7	15	17	16	26	14	1	10
3	7	5	8	14	16	7	29	16	19
4	13	20	20	17	17	14	25	29	10
5	2	8	6	11	30	20	15	5	12
6	12	17	11	9	2	30	22	12	20
7	9	19	12	16	16	25	30	23	19
8	5	9	4	14	23	16	16	30	3

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Utility values are discussed in more detail in Chapter 17.

The part-worths can be used to determine the overall value (utility) that each consumer attaches to a particular type of pizza. For instance, consumer 1's current favorite pizza is the Antonio's brand, which has a thick crust, mozzarella cheese, chunky sauce, and medium-flavored sausage. We can determine consumer 1's utility for this particular type of pizza using the part-worths in Table 15.3. For consumer 1, the part-worths are 2 for thick crust, 6 for mozzarella cheese, 17 for chunky sauce, and 27 for medium-flavored sausage. Thus, consumer 1's utility for the Antonio's brand pizza is $2 + 6 + 17 + 27 = 52$. We can compute consumer 1's utility for a King's brand pizza similarly. The King's brand pizza has a thin crust, a cheese blend, smooth sauce, and mild-flavored sausage. Because the part-worths for consumer 1 are 11 for thin crust, 7 for cheese blend, 3 for smooth sauce, and 26 for mild-flavored sausage, consumer 1's utility for the King's brand pizza is $11 + 7 + 3 + 26 = 47$. In general, each consumer's utility for a particular type of pizza is the sum of the part-worths for the attributes of that type of pizza.

To be successful with its brand, Salem Foods realizes that it must entice consumers in the marketplace to switch from their current favorite brand of pizza to the Salem product. In other words, Salem must design a pizza (choose the type of crust, cheese, sauce, and sausage flavor) that will have the highest utility for enough people to ensure sufficient sales to justify making the product. Assuming the sample of eight consumers in the current study is representative of the marketplace for frozen sausage pizza, we can formulate and solve an integer programming model that can help Salem come up with such a design. In marketing literature, the problem being solved is called the *share of choice* problem.

The decision variables are defined as follows:

l_{ij} = 1 if Salem chooses level i for attribute j; 0 otherwise
y_k = 1 if consumer k chooses the Salem brand; 0 otherwise

The objective is to choose the levels of each attribute that will maximize the number of consumers who prefer the Salem brand pizza. Because the number of consumers who prefer the Salem brand pizza is just the sum of the y_k variables, the objective function is

$$\text{Max } y_1 + y_2 + \cdots + y_8$$

One constraint is needed for each consumer in the sample. To illustrate how the constraints are formulated, let us consider the constraint corresponding to consumer 1. For consumer 1, the utility of a particular type of pizza can be expressed as the sum of the part-worths:

$$\text{Utility for consumer 1} = 11l_{11} + 2l_{21} + 6l_{12} + 7l_{22} + 3l_{13} + 17l_{23} + 26l_{14} + 27l_{24} + 8l_{34}$$

For consumer 1 to prefer the Salem pizza, the utility for the Salem pizza must be greater than the utility for consumer 1's current favorite. Recall that consumer 1's current favorite brand of pizza is Antonio's, with a utility of 52. Thus, consumer 1 will purchase the Salem brand only if the levels of the attributes for the Salem brand are chosen such that

$$11l_{11} + 2l_{21} + 6l_{12} + 7l_{22} + 3l_{13} + 17l_{23} + 26l_{14} + 27l_{24} + 8l_{34} > 52$$

Given the definitions of the y_k decision variables, we want $y_1 = 1$ when the consumer prefers the Salem brand and $y_1 = 0$ when the consumer does not prefer the Salem brand. Thus, we write the constraint for consumer 1 as follows:

$$11l_{11} + 2l_{21} + 6l_{12} + 7l_{22} + 3l_{13} + 17l_{23} + 26l_{14} + 27l_{24} + 8l_{34} \geq 1 + 52y_1$$

With this constraint, y_k cannot equal 1 unless the utility for the Salem design (the left-hand side of the constraint) exceeds the utility for consumer 1's current favorite by at least 1. Because the objective function is to maximize the sum of the y_k variables, the optimization will seek a product design that will allow as many y_k variables as possible to equal 1.

A similar constraint is written for each consumer in the sample. The coefficients for the l_{ij} variables in the utility functions are taken from Table 15.3, and the coefficients for the y_k

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

variables are obtained by computing the overall utility of the consumer's current favorite brand of pizza. The following constraints correspond to the eight consumers in the study:

$$
\begin{aligned}
11l_{11} + 2l_{21} + 6l_{12} + 7l_{22} + 3l_{13} + 17l_{23} + 26l_{14} + 27l_{24} + 8l_{34} &\geq 1 + 52y_1 \\
11l_{11} + 7l_{21} + 15l_{12} + 17l_{22} + 16l_{13} + 26l_{23} + 14l_{14} + 1l_{24} + 10l_{34} &\geq 1 + 58y_2 \\
7l_{11} + 5l_{12} + 8l_{12} + 14l_{22} + 16l_{13} + 7l_{23} + 29l_{14} + 16l_{24} + 19l_{34} &\geq 1 + 66y_3 \\
13l_{11} + 20l_{21} + 20l_{12} + 17l_{22} + 17l_{13} + 14l_{23} + 25l_{14} + 29l_{24} + 10l_{34} &\geq 1 + 83y_4 \\
2l_{11} + 8l_{21} + 6l_{12} + 11l_{22} + 30l_{13} + 20l_{23} + 15l_{14} + 5l_{24} + 12l_{34} &\geq 1 + 58y_5 \\
12l_{11} + 17l_{21} + 11l_{12} + 9l_{22} + 2l_{13} + 30l_{23} + 22l_{14} + 12l_{24} + 20l_{34} &\geq 1 + 70y_6 \\
9l_{11} + 19l_{21} + 12l_{12} + 16l_{22} + 16l_{13} + 25l_{23} + 30l_{14} + 23l_{24} + 19l_{34} &\geq 1 + 79y_7 \\
5l_{11} + 9l_{21} + 4l_{12} + 14l_{22} + 23l_{13} + 16l_{23} + 16l_{14} + 30l_{24} + 3l_{34} &\geq 1 + 59y_8
\end{aligned}
$$

Antonio's brand is the current favorite pizza for consumers 1, 4, 6, 7, and 8. King's brand is the current favorite pizza for consumers 2, 3, and 5.

Four more constraints must be added, one for each attribute. These constraints are necessary to ensure that one and only one level is selected for each attribute. For attribute 1 (crust), we must add the constraint:

$$l_{11} + l_{21} = 1$$

Because l_{11} and l_{21} are both binary variables, this constraint requires that one of the two variables equals one, and the other equals zero. The following three constraints ensure that one and only one level is selected for each of the other three attributes:

$$
\begin{aligned}
l_{12} + l_{22} &= 1 \\
l_{13} + l_{23} &= 1 \\
l_{14} + l_{24} + l_{34} &= 1
\end{aligned}
$$

MODEL *file*
salem

The data, model, and solution for the Salem pizza problem may be found in the file *salem.* The optimal solution to this 17-variable, 12-constraint integer linear program is $l_{11} = l_{22} = l_{23} = l_{14} = 1$ and $y_2 = y_5 = y_6 = y_7 = 1$. The value of the optimal solution is 4, indicating that if Salem makes this type of pizza, it will be preferable to the current favorite for four of the eight consumers. With $l_{21} = l_{22} = l_{23} = l_{14} = 1$, the pizza design that obtains the largest market share for Salem has a thin crust, a cheese blend, a chunky sauce, and mild-flavored sausage. Note also that with $y_2 = y_5 = y_6 = y_7 = 1$, consumers 2, 5, 6, and 7 will prefer the Salem pizza. This information may lead Salem to choose to market this type of pizza.

15.5 Modeling Flexibility Provided by Binary Variables

In Section 15.4, we presented four applications involving binary integer variables. In this section, we continue the discussion of the use of binary integer variables in modeling. First, we show how binary integer variables can be used to model multiple-choice and mutually exclusive constraints. Then we show how binary integer variables can be used to model situations in which k projects out of a set of n projects must be selected, as well as situations in which the acceptance of one project is conditional on the acceptance of another project.

Multiple-Choice and Mutually Exclusive Constraints

Recall the Ice-Cold Refrigerator capital budgeting problem introduced in Section 15.4. The decision variables were defined as follows:

P = 1 if the plant expansion project is accepted; 0 if rejected
W = 1 if the warehouse expansion project is accepted; 0 if rejected
M = 1 if the new machinery project is accepted; 0 if rejected
R = 1 if the new product research project is accepted; 0 if rejected

Suppose that, instead of one warehouse expansion project, the Ice-Cold Refrigerator Company actually has three warehouse expansion projects under consideration. One of the warehouses *must* be expanded because of increasing product demand, but new demand is not sufficient to make expansion of more than one warehouse necessary. The following

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

variable definitions and **multiple-choice constraint** could be incorporated into the previous binary integer linear programming model to reflect this situation. Let:

$W_1 = 1$ if the original warehouse expansion project is accepted; 0 if rejected
$W_2 = 1$ if the second warehouse expansion project is accepted; 0 if rejected
$W_3 = 1$ if the third warehouse expansion project is accepted; 0 if rejected

The multiple-choice constraint reflecting the requirement that exactly one of these projects must be selected is

$$W_1 + W_2 + W_3 = 1$$

If W_1, W_2, and W_3 are allowed to assume only the values 0 or 1, then one and only one of these projects will be selected from among the three choices.

If the requirement that one warehouse must be expanded did not exist, the multiple-choice constraint could be modified as follows:

$$W_1 + W_2 + W_3 \leq 1$$

This modification allows for the case of no warehouse expansion ($W_1 = W_2 = W_3 = 0$) but does not permit more than one warehouse to be expanded. This type of constraint is often called a **mutually exclusive constraint**.

k Out of *n* Alternatives Constraint

An extension of the notion of a multiple-choice constraint can be used to model situations in which *k out of a set of n* projects must be selected—a ***k* out of *n* alternatives constraint**. Suppose that W_1, W_2, W_3, W_4, and W_5 represent five potential warehouse expansion projects and that two of the five projects must be accepted. The constraint that satisfies this new requirement is

$$W_1 + W_2 + W_3 + W_4 + W_5 = 2$$

If no more than two of the projects are to be selected, we would use the following less-than-or-equal-to constraint:

$$W_1 + W_2 + W_3 + W_4 + W_5 \leq 2$$

Again, each of these variables must be restricted to binary values.

Conditional and Corequisite Constraints

Sometimes the acceptance of one project is conditional on the acceptance of another. For example, suppose that for the Ice-Cold Refrigerator Company, the warehouse expansion project was conditional on the plant expansion project. In other words, suppose management will not consider expanding the warehouse unless the plant is expanded. With binary variable P representing plant expansion ($1 =$ expand, $0 =$ do not expand) and W a binary variable representing warehouse expansion ($1 =$ expand, $0 =$ do not expand), a conditional constraint needs to be developed to enforce the requirement the warehouse cannot be expanded unless the plant has been expanded.

When faced with this type of conditional constraint, it is often helpful to construct a **feasibility table**. A feasibility table is a table that lists all possible settings of the relevant binary variables and indicates which settings of these variables are feasible and which are not feasible. In the Ice-Cold Refrigerator case, we have the following feasibility table:

W	*P*	Relationship	Feasible	Rationale
0	0	$W = P$	Yes	We can choose to expand neither.
1	0	$W > P$	No	We cannot expand the warehouse if the plant is not expanded.
0	1	$W < P$	Yes	We can choose not to expand the warehouse, even if we expand the plant.
1	1	$W = P$	Yes	We can expand both the warehouse and the plant.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notice that W is less than or equal to P for the feasible cases and W is greater than P in the infeasible case. Hence, the **conditional constraint** that enforces the restriction is

$$W \leq P$$

Let us consider another situation where the warehouse and plant expansions are dependent on each other. If the warehouse expansion project had to be accepted whenever the plant expansion project was accepted, and vice versa, we would say that we have a **corequisite constraint**. So, if we choose to expand either, the other must be expanded. In this situation, we have the following feasibility table:

W	P	Relationship	Feasible	Rationale
0	0	$W = P$	Yes	We can choose to expand neither.
1	0	$W > P$	No	We cannot expand the warehouse if the plant is not expanded.
0	1	$W < P$	No	If the warehouse is not expanded, we cannot expand the plant.
1	1	$W = P$	Yes	We can expand both the warehouse and the plant.

In this feasibility table, we see that when W and P are set to the same value, the result is feasible, but different settings of W and P are infeasible. Hence, in the corequisite situation, the following constraint enforces the restriction:

$$W = P$$

The constraint forces P and W to take on the same value.

Notes + Comments

1. As in the Ice-Cold Refrigerator examples with conditional and corequisite constraints, many restrictions will involve only two binary variables. Since in the feasibility table, we list all possible cases (settings of the binary variables), there are $2^2 = 4$ cases when there are two variables. Some conditional and corequisite constraints might involve three or more variables. For three variables, there are $2^3 = 8$ cases, for four variables there are $2^4 = 16$ cases. In general, for n variables, there will be 2^n cases. Therefore, for situations involving more than three variables, feasibility tables can become cumbersome.
2. A somewhat natural way to try to model a conditional or corequisite constraint in Excel is to use an IF function. However, since the IF function is a discontinuous function (i.e., a function with a break or jump in the function value), using an IF statement will preclude the use of the LP Simplex option as discussed in Section 15.3. While the nonlinear option in Excel Solver (discussed in Chapter 16) can sometimes find good results even with the use of an IF function, optimality cannot always be guaranteed. Therefore, we recommend you model conditional constraints in a linear way as discussed in Section 15.5.

15.6 Generating Alternatives in Binary Optimization

If alternative optimal solutions exist, it would be good for management to know this because some factors that make one alternative preferred over another might not be included in the model. Also, if the solution is a unique optimal solution, it would be good to know how much worse the second-best solution is than the unique optimal solution. If the second-best solution is very close to optimal, it might be preferred over the true optimal solution because of factors outside the model.

As an example, let us reconsider the Ohio Trust location problem presented in Section 15.4. The solution for the minimum number of principal places of business (PPBs) is three. As shown in Figure 15.11, the solution is to place PPBs in county 7 (Ashland), county 11 (Stark), and county 12 (Geauga). However, suppose when Ohio Trust tries to implement this solution, it is not possible to find a suitable location for a PPB in one of

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

these three counties. Are there other alternative solutions of three counties, or is this a unique optimal solution? By adding a special constraint based on the current solution and then resolving the model, we may answer this question.

The current solution for Ohio Trust can be broken into two sets of variables: those that are set to one and those that are set to zero. Let the set O denote the set of variables set to one and the set Z those that are set to zero. For the Ohio Trust solution, these sets are as follows:

Set O: x_7, x_{11}, x_{12}

Set Z: $x_1, x_2, x_3, x_4, x_5, x_6, x_8, x_9, x_{10}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}$

We may add the following constraint:

(Sum of variables in the set O) $-$ (sum of variables in the set Z) $\leq$ (number of variables in the set O) $- 1$,

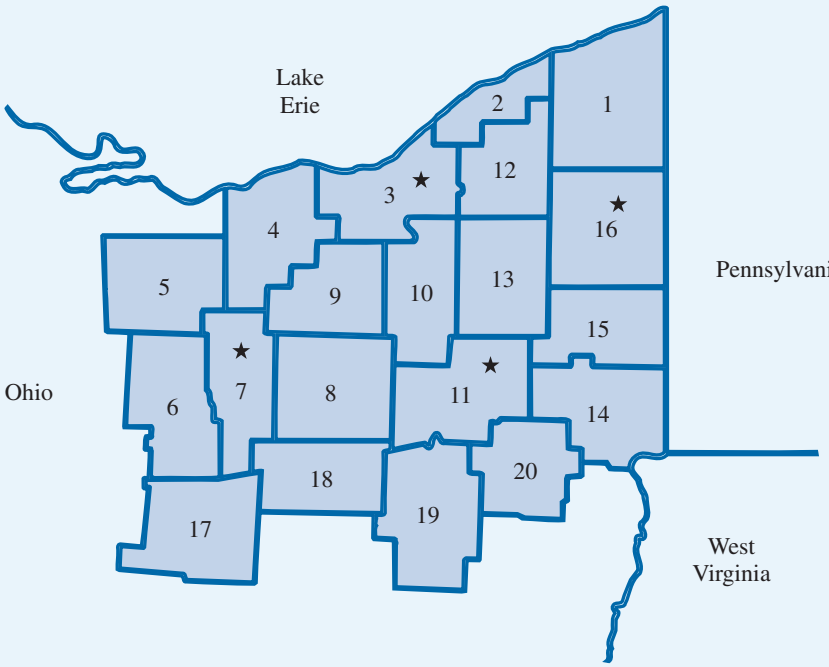
which for our current solution is

$$x_7 + x_{11} + x_{12} - x_1 - x_2 - x_3 - x_4 - x_5 - x_6 - x_8 - x_9 - x_{10} - x_{13} - x_{14} - x_{15} - x_{16} - x_{17} - x_{18} - x_{19} - x_{20} \leq 3 - 1 = 2$$

This constraint has the very special property that it makes the current solution infeasible, but keeps feasible all other solutions that are feasible to the original problem. This constraint will force (at least) one of the variables in set O to change from one to zero or will force (at least) one of the variables in set Z to change from zero to one.

When we append this new constraint to the original model, we obtain the solution displayed in Figure 15.12. Notice that the optimal objective function value has increased to four.

Figure 15.12 A Second-Best Solution to the Ohio Trust Location Problem



Counties

1. Ashtabula	6. Richland	11. Stark	16. Trumbull	★ *A principal place of business should be located in these counties.*
2. Lake	7. Ashland	12. Geauga	17. Knox	
3. Cuyahoga	8. Wayne	13. Portage	18. Holmes	
4. Lorain	9. Medina	14. Columbiana	19. Tuscarawas	
5. Huron	10. Summit	15. Mahoning	20. Carroll	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This tells us that the solution we found in Section 15.4 with objective function value equal to 3 is a unique optimal solution. Any other feasible solution will require four or more PPBs to cover the entire 20-county region. So, if for any of the three counties in the original solution we cannot find a suitable location for a PPB, the next-best solution will require PPBs in four counties and the solution in Figure 15.10 is a second-best alternative. Note that if the optimal objective functions of the new problem with constraint added had been 3, we would have found an alternative optimal solution.

We can summarize the procedure for finding an alternative solution as follows:

Step 1. Solve the original problem

Step 2. Create two sets:

O = the set of variables equal to one in Step 1

Z = the set of variables equal to zero in Step 1

Step 3. Add the following constraint to the original problem, and solve

$$\begin{aligned}&(\text{Sum of variables in the set O}) - (\text{sum of variables in the set Z})\\ &\quad \leq (\text{number of variables in the set O}) - 1\end{aligned} \qquad \textbf{(15.1)}$$

If the objective function value in Step 3 is equal to the objective function value of Step 1, we have found an alternative optimal solution. If the objective function value of Step 3 is inferior to that of Step 1, we have found a next-best solution.

Notes + Comments

1. The procedure just described can be applied iteratively. In other words, we can take the second-best solution found and create the equation (15.1) based on that solution to find the next-best solution. Note that we leave all previous constraints in the problem, including the first constraint based on equation (15.1). The resulting solution could be a third-best solution or an alternative second-best solution. It turns out that there are numerous second-best solutions to the Ohio Trust problem using four PPBs.
2. Applying equation (15.1) iteratively and finding that the objective function value does not deteriorate, generates an alternative optimal solution. In fact applying equation (15.1) iteratively until the objective function changes ensures you have found all alternative optima.

Summary

In this chapter, we introduced the important extension of linear programming referred to as integer linear programming. The only difference between the integer linear programming problems discussed in this chapter and the linear programming problems studied in the previous chapter is that one or more of the variables must be an integer. If all variables must be integer, we have an all-integer linear program. If some, but not necessarily all, variables must be an integer, we have a mixed-integer linear program. Most integer programming applications involve binary variables.

Studying integer linear programming is important for two major reasons. First, integer linear programming may be helpful when fractional values for the variables are not permitted. Rounding a linear programming solution may not provide an optimal integer solution; methods for finding optimal integer solutions are needed when the economic consequences of rounding are substantial. A second reason for studying integer linear programming is the increased modeling flexibility provided through the use of binary variables. We showed how binary variables could be used to model important managerial considerations in capital budgeting, fixed cost, facility location, and product design/market share applications. We showed how to generate second-best solutions or alternative optima if they exist by adding a constraint based on those solutions. This is important for providing alternatives for management.

The number of applications of integer linear programming continues to grow rapidly, partly because of the availability of good integer linear programming software packages.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

As researchers develop solution procedures capable of solving larger integer linear programs and as computer speed increases, a continuation of the growth of integer programming applications is expected.

Glossary

All-integer linear program An integer linear program in which all variables are required to be integers.
Binary integer linear program An all-integer or mixed-integer linear program in which the integer variables are permitted to assume only the values 0 or 1. Also called *binary integer program*.
Capital budgeting problem A binary integer programming problem that involves choosing which possible projects or activities provide the best investment return.
Conditional constraint A constraint involving binary variables that does not allow certain variables to equal one unless certain other variables are equal to one.
Conjoint analysis A market research technique that can be used to learn how prospective buyers of a product value the product's attributes.
Convex hull The smallest intersection of linear inequalities that contain a certain set of points.
Corequisite constraint A constraint requiring that two binary variables be equal and that they are both either in or out of the solution.
Feasibility table A table that is useful in modeling conditional and corequisite constraints with binary variables. The table lists all possible settings of the relevant binary variables and indicates which settings of these variables are feasible and which are not feasible.
Fixed-cost problem A binary mixed-integer programming problem in which the binary variables represent whether an activity, such as a production run, is undertaken (variable = 1) or not (variable = 0).
Integer linear program A linear program with the additional requirement that one or more of the variables must be an integer.
***k* out of *n* alternatives constraint** An extension of the multiple-choice constraint that requires that the sum of n binary variables equals k.
Location problem A binary integer programming problem in which the objective is to select the best locations to meet a stated objective. Variations of this problem (see the bank location problem in Section 15.4) are known as *covering problems*.
LP relaxation The linear program that results from dropping the integer requirements for the variables in an integer linear program.
Mixed-integer linear program An integer linear program in which some, but not necessarily all, variables are required to be integers.
Multiple-choice constraint A constraint requiring that the sum of two or more binary variables equals one. Thus, any feasible solution makes a choice of which variable to set equal to one.
Mutually exclusive constraint A constraint requiring that the sum of two or more binary variables be less than or equal to one. Thus, if one of the variables equals one, the others must equal zero. However, all variables could equal zero.
Part-worth The utility value that a consumer attaches to each level of each attribute in a conjoint analysis model.
Product design and market share optimization problem Sometimes called the *share of choice problem*, the choice of a product design that maximizes the number of consumers that prefer it.

Problems

1. **Machine Tool Production Planning.** King City Inc. manufactures machine tools. The production planner who oversees the production of two of King City's machines needs to determine how many of each to produce this month. The two machines, TopLathe

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

and BigPress, each require a certain common component. Each TopLathe requires 10 of these components and each BigPress requires 7. Only 49 components are available this month. The sales department requires that the total number of machines produced in a month must be at least 5 (the number TopLathes plus the number BigPresses must be at least 5). The profit for a TopLathe is $50,000 and $34,000 for a BigPress. **LO 1, 4, 5**

a. Assuming that adequate labor and all other resources are available, formulate an integer programming model to determine how many of each product King City should produce to maximize profit.
b. Solve the model formulated in part (a) without integer requirements. What is the optimal profit? What are the optimal values for TopLathe and BigPress?
c. Round the TopLathe and BigPress values found in part (b). Is the solution feasible? Why?
d. Truncate the TopLathe and BigPress values found in part (b) (drop the fractional part of each value). Is the solution feasible? Why?
e. Add integer requirements to the model you constructed in part (b). What is the optimal profit and what are the optimal number of TopLathes and BigPresses?

2. **Nurse Scheduling.** Hospital administrators must schedule nurses so that the hospital's patients are provided adequate care. At the same time, careful attention must be paid to keeping costs down. From historical records, administrators can project the minimum number of nurses required to be on hand for various times of day and days of the week. The objective is to find the minimum total number of nurses required to provide adequate care.

Nurses start work at the beginning of one of the four-hour shifts given below (except for shift 6) and work for 8 consecutive hours. Hence, possible start times are the start of shifts 1 through 5. Also, assume that the projected required number of nurses factors in time for each nurse to have a meal break.

Formulate and solve the nurse scheduling problem as an integer program for one day for the data given below. **LO 1, 4, 5**

Hint: Note that exceeding the minimum number of needed nurses in each shift is acceptable so long as the total number of nurses over all shifts is minimized.

DATA*file*
nurseschedule

Shift	Time	Minimum Number of Nurses Needed
1	12:00 a.m.–4:00 a.m.	10
2	4:00 a.m.–8:00 a.m.	24
3	8:00 a.m.–12:00 p.m.	18
4	12:00 p.m.–4:00 p.m.	10
5	4:00 p.m.–8:00 p.m.	23
6	8:00 p.m.–12:00 a.m.	17

3. **Minimizing Scrap.** STAR Co. provides paper to smaller companies with volumes that are not large enough to warrant dealing directly with the paper mill. STAR receives 100-feet-wide paper rolls from the mill and cuts the rolls into smaller rolls of widths 12, 15, and 30 feet. The demands for these widths vary from week to week. The following cutting patterns have been established:

DATA*file*
scrap

Pattern Number	12-ft	15-ft	30-ft	Trim Loss (ft)
1	0	6	0	10
2	0	0	3	10
3	8	0	0	4
4	2	1	2	1
5	7	1	0	1

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Trim loss is the leftover paper from a pattern (e.g., for pattern 4, $2(12) + 1(15) + 2(30) = 99$ feet used results in $100 - 99 = 1$ foot of trim loss). Orders in hand for the coming week are 5,670 12-foot rolls, 1,680 15-foot rolls, and 3,350 30-foot rolls. Any of the three types of rolls produced in excess of the orders in hand will be sold on the open market at the selling price. No inventory is held. **LO 1, 4, 5**

a. Formulate an integer programming model that will determine how many 100-foot rolls to cut into each of the five patterns in order to minimize trim loss.
b. Solve the model formulated in part (a). What is the minimal amount of trim loss? How many of each pattern should be used and how many of each type of roll will be sold on the open market?

4. **Real Estate Project Selection.** Brooks Development Corporation (BDC) faces the following capital budgeting decision. Six real estate projects are available for investment. The net present value and expenditures required for each project (in millions of dollars) are as follows:

DATA*file* brooks

Project	1	2	3	4	5	6
Net Present Value ($ millions)	$15	$5	$13	$14	$20	$9
Expenditure Required ($ millions)	$90	$34	$81	$70	$114	$50

There are conditions that limit the investment alternatives:

- At least two of projects 1, 3, 5, and 6 must be undertaken.
- If either project 3 or 5 is undertaken, they must both be undertaken.
- Project 4 cannot be undertaken unless both projects 1 and 3 also are undertaken.

The budget for this investment period is $220 million. **LO 2, 4, 5, 7**

a. Formulate a binary integer program that will enable BDC to find the projects to invest in to maximize net present value, while satisfying all project restrictions and not exceeding the budget.
b. Solve the model formulated in part (a). What is the optimal net present value? Which projects will be undertaken? How much of the budget is unused?

5. **Investment Net Present Value.** Spencer Enterprises is attempting to choose among a series of new investment alternatives. The potential investment alternatives, the net present value of the future stream of returns, the capital requirements, and the available capital funds over the next three years are summarized as follows. At most one of the warehouse expansion alternatives can be chosen. **LO 2, 4, 5, 7**

DATA*file* investmentnpv

		Capital Requirements ($)		
Alternative	Net Present Value ($)	Year 1	Year 2	Year 3
Limited warehouse expansion	4,000	1,000	1,000	4,000
Extensive warehouse expansion	6,000	2,500	3,500	3,500
Test market new product	10,500	6,000	4,000	5,000
Advertising campaign	4,000	2,000	1,500	1,800
Basic research	8,000	5,000	1,000	4,000
Purchase new equipment	3,000	1,000	500	900
Capital funds available		9,000	11,000	12,000

a. Develop and solve an integer programming model for maximizing the net present value.
b. Suppose that if the Advertising campaign is chosen, then the Test market new product alternative must be chosen. Modify your model from part (a) to take this into account and solve.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Suppose the Extensive warehouse expansion must be chosen. Modify your formulation from part (b) to reflect this new situation and solve. What is the impact of this constraint on the optimal NPV in comparison to the optimal NPV from part (b)?

6. **Component Ordering.** Morgan Inc. is planning the purchase of one of the component parts it needs for its finished product. The anticipated demands for the component for the next 12 periods are shown in the following table. The cost to order the component (labor, shipping, and paperwork) is a fixed cost of $150 per order. The cost to hold these components in inventory is $1 per component per period. The price of the component is expected to remain stable at $12 per unit for the next 12 periods, and no quantity discounts are available. The maximum order size is 1,000 units. **LO 3, 4, 5**

DATA *file* morgan

Period	1	2	3	4	5	6	7	8	9	10	11	12
Demand	20	20	30	40	140	360	500	540	460	80	0	20

a. Formulate a model to minimize the total cost of satisfying Morgan Inc.'s demand for this component.
b. Build a spreadsheet model for the model formulated in part (a).
c. What is the optimal cost? How many orders are placed?

7. **Locating Police Substations.** Grave City is considering the relocation of several police substations to obtain better enforcement in high-crime areas. The locations under consideration together with the areas that can be covered from these locations are given in the following table. A one in the table below indicates that the substation covers the area and a 0 indicates it does not cover the area. **LO 2, 4, 5**

DATA *file* policesub

	Substation A	Substation B	Substation C	Substation D	Substation E	Substation F	Substation G
Area 1	1	1	1	0	0	0	1
Area 2	0	1	0	1	0	0	0
Area 3	0	0	1	0	1	0	0
Area 4	0	0	0	1	1	1	0
Area 5	1	1	1	1	0	1	1
Area 6	0	0	0	0	1	1	1
Area 7	1	1	0	0	0	0	1

a. Formulate an integer programming model that could be used to find the minimum number of locations necessary to provide coverage to all areas.
b. Build a spreadsheet model for this problem.
c. Solve the model. Which locations are chosen?

8. **Multi-product Production Planning.** Hart Manufacturing makes three products. Each product requires manufacturing operations in three departments: A, B, and C. The labor-hour requirements, by department, are as follows:

DATA *file* mppp

Department	Product 1	Product 2	Product 3
A	1.50	3.00	2.00
B	2.00	1.00	2.50
C	0.25	0.25	0.25

During the next production period the labor-hours available are 450 in department A, 350 in department B, and 50 in department C. The profit contributions per unit are $25 for product 1, $28 for product 2, and $30 for product 3. **LO 3, 4, 5, 7**

a. Formulate a linear programming model for maximizing total profit contribution.
b. Solve the linear program formulated in part (a). How much of each product should be produced, and what is the projected total profit contribution?
c. After evaluating the solution obtained in part (b), one of the production supervisors noted that production setup costs had not been taken into account. The supervisor noted that setup costs are $400 for product 1, $550 for product 2, and $600 for

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

product 3. If the solution developed in part (b) is to be used, what is the total profit contribution after taking into account the setup costs?

d. Management realized that the optimal product mix, taking setup costs into account, might be different from the one recommended in part (b). Formulate a mixed-integer linear program that takes setup costs provided in part (c) into account. Management also stated that we should not consider making more than 175 units of product 1, 150 units of product 2, or 140 units of product 3.

e. Solve the mixed-integer linear program formulated in part (d). How much of each product should be produced and what is the projected total profit contribution? Compare this profit contribution to that obtained in part (c).

9. **Carrier Selection.** Offhaus Manufacturing produces office supplies but outsources the delivery of its products to third-party carriers. Offhaus ships to 20 cities from its Dayton, Ohio, manufacturing facility and has asked a variety of carriers to bid on its business. Seven carriers have responded with bids. The resulting bids (in dollars per truckload) are shown in the table. For example, the table shows that carrier 1 bid on the business to cities 11 to 20. The right side of the table provides the number of truckloads scheduled for each destination in the next quarter.

Bid $/truckload	Carrier 1	Carrier 2	Carrier 3	Carrier 4	Carrier 5	Carrier 6	Carrier 7	Destination	Demand (truckloads)
City 1					2,188	1,666	1,790	City 1	30
City 2		1,453			2,602	1,767		City 2	10
City 3		1,534			2,283	1,857	1,870	City 3	20
City 4		1,687			2,617	1,738		City 4	40
City 5		1,523			2,239	1,771	1,855	City 5	10
City 6		1,521			1,571		1,545	City 6	10
City 7		2,100		1,922	1,938		2,050	City 7	12
City 8		1,800		1,432	1,416		1,739	City 8	25
City 9		1,134		1,233	1,181		1,150	City 9	25
City 10		672		610	669		678	City 10	33
City 11	724		723	627	657		706	City 11	11
City 12	766		766	721	682		733	City 12	29
City 13	741		745		682		733	City 13	12
City 14	815	800	828		745		832	City 14	24
City 15	904		880		891		914	City 15	10
City 16	958		933		891		914	City 16	10
City 17	925		929		937		984	City 17	23
City 18	892		869	822	829		864	City 18	25
City 19	927		969		967		1,008	City 19	12
City 20	963		938		955		995	City 20	10
No. of Bids	10	10	10	7	20	5	18		

DATA *file*
offhaus

Because dealing with too many carriers can be cumbersome, Offhaus would like to limit the number of carriers it uses to three. Also, for customer relationship reasons Offhaus wants each city to be assigned to only one carrier (i.e., no splitting of the demand to a given city across carriers). **LO 2, 4, 5, 6**

a. Develop a model that will yield the three selected carriers and the city-carrier assignments that minimize the cost of shipping. Solve the model and report the solution.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Offhaus is not sure whether three is the correct number of carriers to select. Run the model you developed in part (a) for allowable carriers varying from one to seven. Based on results, how many carriers would you recommend and why?

10. **Manufacturing Plant Location.** The Martin-Beck Company operates a plant in St. Louis with an annual capacity of 30,000 units. Product is shipped to regional distribution centers located in Boston, Atlanta, and Houston. Because of an anticipated increase in demand, Martin-Beck plans to increase capacity by constructing a new plant in one or more of the following cities: Detroit, Toledo, Denver, or Kansas City. The estimated annual fixed cost and the annual capacity for the four proposed plants are as follows:

DATA*file*
plant_location

Proposed Plant	Annual Fixed Cost ($)	Annual Capacity
Detroit	175,000	10,000
Toledo	300,000	20,000
Denver	375,000	30,000
Kansas City	500,000	40,000

The company's long-range planning group developed forecasts of the anticipated annual demand at the distribution centers as follows. **LO 3, 4, 5, 8**

Distribution Center	Annual Demand
Boston	30,000
Atlanta	20,000
Houston	20,000

The shipping cost per unit (in dollars) from each plant to each distribution center is as follows:

	Distribution Centers		
Plant Site	Boston	Atlanta	Houston
Detroit	5	2	3
Toledo	4	3	4
Denver	9	7	5
Kansas City	10	4	2
St. Louis	8	4	3

a. Formulate a mixed-integer programming model that could be used to help Martin-Beck determine which new plant or plants to open in order to satisfy anticipated demand.
b. Solve the model you formulated in part (a). What is the optimal cost? What is the optimal set of plants to open?
c. Using equation (15.1), find a second-best solution. What is the increase in cost versus the best solution from part (b)?

11. **Cloud Services Capacity Planning.** Galaxy Cloud Services operates several data centers across the United States containing servers that store and process the data on the Internet. Suppose that Galaxy Cloud Services currently has five outdated data centers: one each in Michigan, Ohio, and California and two in New York. Management is considering increasing the capacity of these data centers to keep up with increasing demand. Each data center contains servers that are dedicated to Secure data and to

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Super Secure data. The cost to update each data center and the resulting increase in server capacity for each type of server are as follows:

DATA *file* cloudservers

		Increase in Capacity	
Data Center	Cost ($ millions)	Secure Servers	Super Secure Servers
Michigan	2.5	50	30
New York 1	3.5	80	40
New York 2	3.5	40	80
Ohio	4.0	90	60
California	2.0	20	30

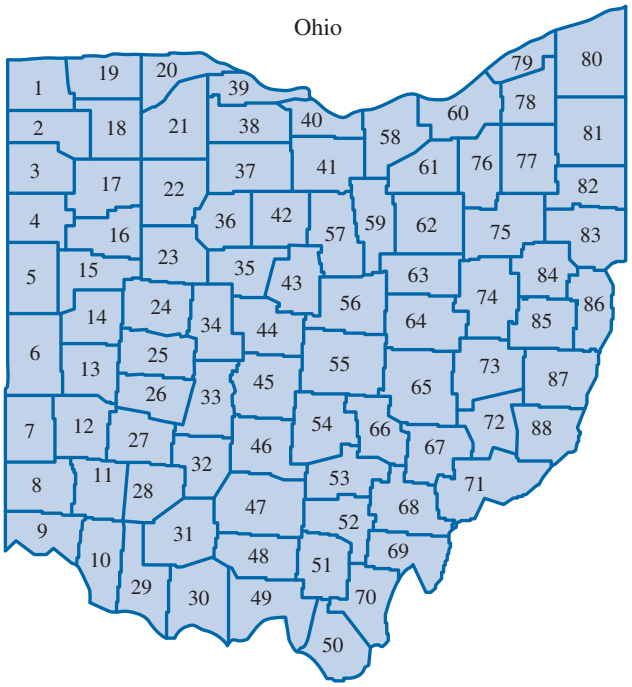
The projected needs are for a total increase in capacity of 90 Secure servers and 90 Super Secure servers. Management wants to determine which data centers to update to meet projected needs and, at the same time, minimize the total cost of the added capacity. **LO 2, 4, 5**

a. Formulate a binary integer programming model that could be used to determine the optimal solution to the capacity increase question facing management.
b. Solve the model formulated in part (a) to provide a recommendation for management.

12. **Bank Location Planning.** CHB, Inc., a bank holding company, is evaluating the potential for expanding into the State of Ohio. State law permits establishing branches in any county that is adjacent to a county in which a PPB (principal place of business) is located. The following map shows the State of Ohio. The file *chb* contains an adjacency matrix with a one in the *i*th row and *j*th column indicating that the counties represented by the *i*th row and the *j*th column share a border. A zero indicates that the two counties do not share a border.

Formulate and solve a linear binary model that will tell CHB the minimum number of PPBs required and their location in order to allow CHB to put a branch in every county in Ohio. What is the optimal value of the objective function? **LO 2, 4**

DATA *file* chb



13. **Bank Location Planning (Revisited).** Consider again Problem 12. Use equation (15.1) to determine whether your solution to Problem 12 is the unique optimal solution. If

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the optimal solution is not unique, use equation (15.1) iteratively to find all alternative optimal solutions. How many are there? **LO 8**

14. **Population Reach.** Consider again the CHB, Inc. problem described in Problem 12. Suppose only a limited number of PPBs can be placed. CHB would like to place this limited number of PPBs in counties so that the allowable branches can reach the maximum possible population. The file *chbpop* contains the county adjacency matrix described in Problem 12 as well as the population of each county. **LO 2, 4, 5, 6, 7**

DATA*file* chbpop

    a. Assume that only a fixed number of PPBs, denoted by k, can be established. Formulate a linear binary integer program that will tell CHB, Inc. where to locate the fixed number of PPBs in order to maximize the population reached. (*Hint:* Review the Ohio Trust formulation in Section 15.4. Introduce a binary variable y_i such that $y_i = 1$ if county i can be reached by a PPB (because there is a PPB in county i or in an adjacent county to county i), and $y_i = 0$ otherwise.)
    b. Suppose that two PPBs can be established. Where should they be located to maximize the population served?
    c. Solve your model from part (a) for an allowable number of PPBs ranging from 1 to 10. In other words, solve the model 10 times, k set to 1, 2, . . . , 10. Record the population reached for each value of k. Graph the results by plotting the population reached versus the number of PPBs allowed. Based on their cost calculations, CHB considers an additional PPB to be fiscally prudent only if it increases the population reached by at least 500,000 people. Based on this graph, how many PPBs do you recommend to implement?

15. **Bank Teller Scheduling.** The Northside Bank is working to develop an efficient work schedule for full-time and part-time tellers. The schedule must provide for efficient operation of the bank, including adequate customer service, employee breaks, and so on. On Fridays, the bank is open from 9:00 a.m. to 7:00 p.m. The number of tellers necessary to provide adequate customer service during each hour of operation is summarized as follows:

DATA*file* bank_tellers

Time	Number of Tellers	Time	Number of Tellers
9:00 a.m.–10:00 a.m.	6	2:00 p.m.–3:00 p.m.	6
10:00 a.m.–11:00 a.m.	4	3:00 p.m.–4:00 p.m.	4
11:00 a.m.–Noon	8	4:00 p.m.–5:00 p.m.	7
Noon–1:00 p.m.	10	5:00 p.m.–6:00 p.m.	6
1:00 p.m.–2:00 p.m.	9	6:00 p.m.–7:00 p.m.	6

Each full-time employee starts on the hour and works a 4-hour shift, followed by a 1-hour break and then a 3-hour shift. Part-time employees work one 4-hour shift beginning on the hour. Considering salary and fringe benefits, full-time employees cost the bank $15 per hour ($105 a day), and part-time employees cost the bank $8 per hour ($32 per day). **LO 1, 4, 5**

    a. Formulate an integer programming model that can be used to develop a schedule that will satisfy customer service needs at a minimum employee cost. (*Hint:* Let x_i = number of full-time employees coming on duty at the beginning of hour i and y_i = number of part-time employees coming on duty at the beginning of hour i.)
    b. Solve the LP relaxation of your model in part (a).
    c. Solve your model in part (a) for the optimal schedule of tellers. Comment on the solution.
    d. After reviewing the solution to part (c), the bank manager realized that some additional requirements must be specified. Specifically, the bank manager wants to ensure that one full-time employee is on duty at all times and that there is a staff of at least five full-time employees. Revise your model to incorporate these additional requirements, and solve for the optimal solution.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

16. **Product Design.** Burnside Marketing Research conducted a study for Barker Foods on several formulations for a new dry cereal. Three attributes were found to be most influential in determining which cereal had the best taste: ratio of wheat to corn in the cereal flake, type of sweetener (sugar, honey, or artificial), and the presence or absence of flavor bits. Seven children participated in taste tests and provided the following part-worths for the attributes (see Section 15.4 for a discussion of part-worths). **LO 2, 4, 5, 7**

DATA*file* burnside

	Wheat/Corn		Sweetener			Flavor Bits	
Child	Low	High	Sugar	Honey	Artificial	Present	Absent
1	15	35	30	40	25	15	9
2	30	20	40	35	35	8	11
3	40	25	20	40	10	7	14
4	35	30	25	20	30	15	18
5	25	40	40	20	35	18	14
6	20	25	20	35	30	9	16
7	30	15	25	40	40	20	11

a. Suppose that the overall utility (sum of part-worths) of the current favorite cereal is 75 for each child. What product design will maximize the number of children in the sample who prefer the new dry cereal? Note that a child will prefer the new dry cereal only if its overall utility is at least 1 part-worth larger than the utility of their current preferred cereal.

b. Assume that the overall utility of the current favorite cereal for children 1 to 4 is 70, and the overall utility of the current favorite cereal for children 5 to 7 is 80. What product design will maximize the number of children in the sample who prefer the new dry cereal? Note that a child will prefer the new dry cereal only if its overall utility is at least 1 part-worth larger than the utility of their current preferred cereal.

17. **Security System Design.** The Bayside Art Gallery is considering installing a video camera security system to reduce its insurance premiums. A diagram of Bayside's eight exhibition rooms is shown in the figure in the next page; the openings between the rooms are numbered 1 to 13. A security firm proposed that two-way cameras be installed at some room openings. Each camera has the ability to monitor the two rooms between which the camera is located. For example, if a camera were located at opening number 4, rooms 1 and 4 would be covered; if a camera were located at opening 11, rooms 7 and 8 would be covered; and so on. Management decided not to locate a camera system at the entrance to the display rooms. The objective is to provide security coverage for all eight rooms using the minimum number of two-way cameras. **LO 2, 4, 5**

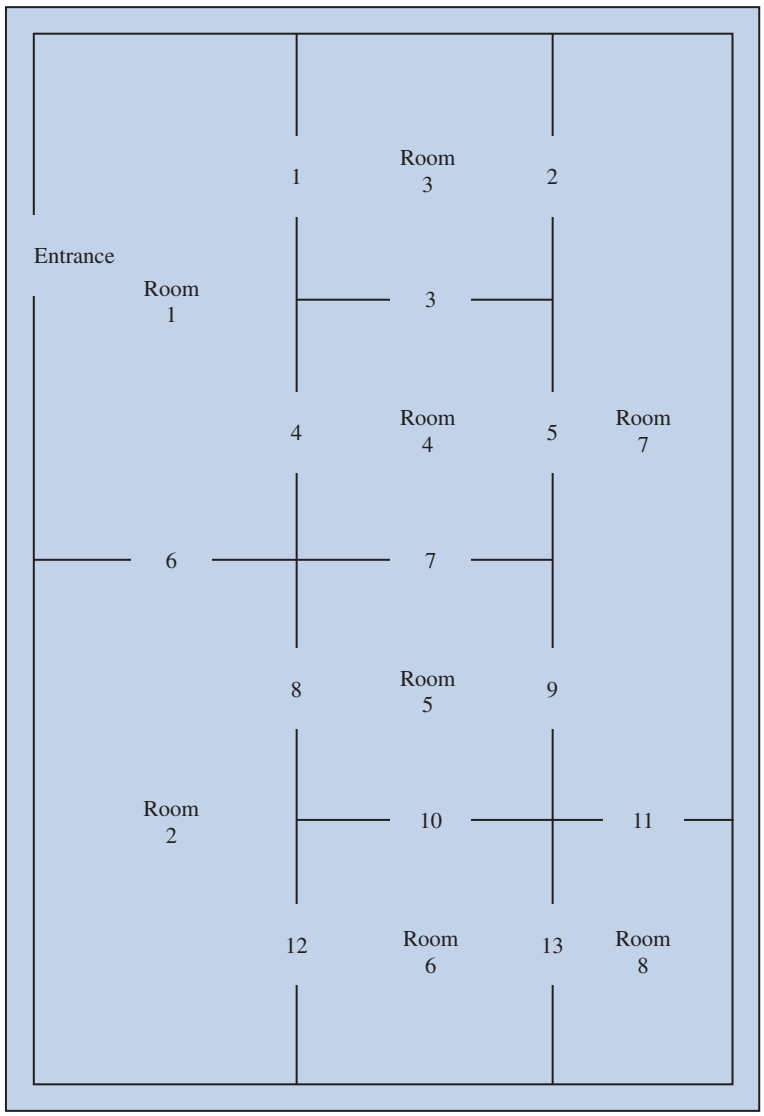
a. Formulate a binary integer linear programming model that will enable Bayside's management to determine the locations for the camera systems.

b. Solve the model formulated in part (a) to determine how many two-way cameras to purchase and where they should be located.

c. Suppose that management wants to provide additional security coverage for room 7. Specifically, management wants room 7 to be covered by two cameras. How would the model you formulated in part (a) have to change to accommodate this policy restriction?

d. With the policy restriction specified in part (c), determine how many two-way camera systems will need to be purchased and where they should be located.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



18. **Fabrics Make Versus Buy.** The Calhoun Textile Mill is in the process of deciding on a production schedule. It wishes to know how to weave the various fabrics it will produce during the coming quarter. The sales department has confirmed orders for each of the 15 fabrics produced by Calhoun. These demands are given in the following table. Also given in this table is the variable cost for each fabric. The mill operates continuously during the quarter: 13 weeks, 7 days a week, and 24 hours a day.

    There are two types of looms: dobby and regular. Dobby looms can be used to make all fabrics and are the only looms that can weave certain fabrics, such as plaids. The rate of production for each fabric on each type of loom is also given in the table. Note that if the production rate is zero, the fabric cannot be woven on that type of loom. Also, if a fabric can be woven on each type of loom, then the production rates are equal. Calhoun has 90 regular looms and 15 dobby looms. For this problem, assume that the time requirement to change over a loom from one fabric to another is negligible.

    In addition to producing the fabric using dobby and regular looms, Calhoun has the option to buy each fabric on the market. The market cost per yard for each fabric is given in the table. Management wants to make sure that each fabric is sourced using exactly one of the three alternatives. That is, for a given fabric, its demand must be

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

satisfied completely by utilizing exclusively only dobby looms, only regular looms, or only being bought on the market.

Management would like to know how to allocate the looms to the fabrics and which fabrics to buy on the market so as to minimize the cost of meeting demand. **LO 3, 4, 5, 7**

DATA *file* calhoun_fabrics

Fabric	Demand (yd)	Dobby (yd/hr)	Regular (yd/hr)	Mill Cost ($/yd)	Market Cost ($/yd)
1	16,500	4.653	0.000	0.6573	0.80
2	52,000	4.653	0.000	0.5550	0.70
3	45,000	4.653	0.000	0.6550	0.85
4	22,000	4.653	0.000	0.5542	0.70
5	76,500	5.194	5.194	0.6097	0.75
6	110,000	3.809	3.809	0.6153	0.75
7	122,000	4.185	4.185	0.6477	0.80
8	62,000	5.232	5.232	0.4880	0.60
9	7,500	5.232	5.232	0.5029	0.70
10	69,000	5.232	5.232	0.4351	0.60
11	70,000	3.733	3.733	0.6417	0.80
12	82,000	4.185	4.185	0.5675	0.75
13	10,000	4.439	4.439	0.4952	0.65
14	380,000	5.232	5.232	0.3128	0.45
15	62,000	4.185	4.185	0.5029	0.70

19. **Subassembly Make versus Buy.** Roedel Electronics produces tablet computer accessories, including integrated keyboard tablet stands that connect a keyboard to a tablet device and holds the device at a preferred angle for easy viewing and typing. Roedel produces two sizes of integrated keyboard tablet stands, small and large. Each size uses the same keyboard attachment, but the stand consists of two different pieces, a top flap and a vertical stand that differ by size. Thus, a completed integrated keyboard tablet stand consists of three subassemblies that are manufactured by Roedel: a keyboard, a top flap, and a vertical stand.

Roedel's sales forecast indicates that 7,000 small integrated keyboard tablet stands and 5,000 large integrated keyboard tablet stands will be needed to satisfy demand during the upcoming Christmas season. Because only 500 hours of in-house manufacturing time are available, Roedel is considering purchasing some, or all, of the subassemblies from outside suppliers. If Roedel manufactures a subassembly in-house, it incurs a fixed setup cost as well as a variable manufacturing cost. The following table shows the setup cost, the manufacturing time per subassembly, the manufacturing cost per subassembly, and the cost to purchase each of the subassemblies from an outside supplier. **LO 3, 4, 5, 7**

DATA *file* subassembly

Subassembly	Setup Cost ($)	Manufacturing Time per Unit (minutes)	Manufacturing Cost per Unit ($)	Purchase Cost per Unit ($)
Keyboard	1,000	0.9	0.40	0.65
Small top flap	1,200	2.2	2.90	3.45
Large top flap	1,900	3.0	3.15	3.70
Small vertical stand	1,500	0.8	0.30	0.50
Large vertical stand	1,500	1.0	0.55	0.70

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. Determine how many units of each subassembly Roedel should manufacture and how many units of each subassembly Roedel should purchase. What is the total manufacturing and purchase cost associated with your recommendation?
b. Suppose Roedel is considering purchasing new machinery to produce large top flaps. For the new machinery, the setup cost is $3,000; the manufacturing time is 2.5 minutes per unit, and the manufacturing cost is $2.60 per unit. Assuming that the new machinery is purchased, determine how many units of each subassembly Roedel should manufacture and how many units of each subassembly Roedel should purchase. What is the total manufacturing and purchase cost associated with your recommendation? Do you think the new machinery should be purchased? Explain.

20. **Television Show Planning.** John White is the program scheduling manager for the television channel CCFO. John would like to plan the schedule of television shows for next Wednesday evening.

The table below lists nine shows under consideration. John must select exactly five of these shows for the period from 8:00 p.m. to 10:30 p.m. next Wednesday evening. For each television show, the estimated advertising revenue (in $ millions) is provided. Furthermore, each show has been categorized into one or more of the categories "Public Interest," "Violent," "Comedy," and "Drama." In the following table, a 1 indicates that the show is in the corresponding category and a 0 indicates it is not.

DATA *file*
tvschedule

Show	Revenue ($ millions)	Public Interest	Violent	Comedy	Drama
Sam's Place	$6	0	0	1	1
Texas Oil	$10	0	1	0	1
Cincinnati Law	$9	1	0	0	1
Jarred	$4	0	1	0	1
Bob & Mary	$5	0	0	1	0
Chainsaw	$2	0	1	0	0
Loving Life	$6	1	0	0	1
Islanders	$7	0	0	1	0
Urban Sprawl	$8	1	0	0	0

John would like to determine a revenue-maximizing schedule of television shows for next Wednesday evening. However, John must be mindful of the following considerations. **LO 2, 4, 5, 7**

- The schedule must include at least as many shows that are categorized as public interest as shows that are categorized as violent.
- If John schedules "Loving Life," then John must also schedule either "Jarred" or "Cincinnati Law" (or both).
- John cannot schedule both "Loving Life" and "Urban Sprawl."
- If John schedules more than one show in the "Violent" category, John will lose an estimated $4 million in advertising revenues from family-oriented sponsors.

a. Formulate a binary integer program that models the decisions John faces.
b. Solve the model formulated in part (a). What is the optimal revenue?

21. **Service Facility Location.** East Coast Trucking provides service from Boston to Miami using regional offices located in Boston, New York, Philadelphia, Baltimore, Washington, Richmond, Raleigh, Florence, Savannah, Jacksonville, Tampa, and Miami. The number of miles between the regional offices is provided in the following table:

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

	Boston	New York	Philadelphia	Baltimore	Washington	Richmond	Raleigh	Florence	Savannah	Jacksonville	Tampa	Miami
Boston	0	211	320	424	459	565	713	884	1,056	1,196	1,399	1,669
New York	211	0	109	213	248	354	502	673	845	985	1,188	1,458
Philadelphia	320	109	0	104	139	245	393	564	736	876	1,079	1,349
Baltimore	424	213	104	0	35	141	289	460	632	772	975	1,245
Washington	459	248	139	35	0	106	254	425	597	737	940	1,210
Richmond	565	354	245	141	106	0	148	319	491	631	834	1,104
Raleigh	713	502	393	289	254	148	0	171	343	483	686	956
Florence	884	673	564	460	425	319	171	0	172	312	515	785
Savannah	1,056	845	736	632	597	491	343	172	0	140	343	613
Jacksonville	1,196	985	876	772	737	631	483	312	140	0	203	473
Tampa	1,399	1,188	1,079	975	940	834	686	515	343	203	0	270
Miami	1,669	1,458	1,349	1,245	1,210	1,104	956	785	613	473	270	0

DATA *file*
eastcoast

The company's expansion plans involve constructing service facilities in some of the cities where regional offices are located. Each regional office must be within 400 miles of a service facility. For instance, if a service facility is constructed in Richmond, it can provide service to regional offices located in New York, Philadelphia, Baltimore, Washington, Richmond, Raleigh, and Florence. Management would like to determine the minimum number of service facilities needed and where they should be located. **LO 2, 4, 5, 7**

a. Formulate an integer linear program that can be used to determine the minimum number of service facilities needed and their locations.
b. Solve the integer linear program formulated in part (a). How many service facilities are required, and where should they be located?
c. Suppose that each service facility can provide service only to regional offices within 300 miles. Re-solve the integer linear program with the 300-mile requirement. How many service facilities are required and where should they be located?

22. **Mutual Fund Portfolio Planning.** Dave has $100,000 to invest in 10 mutual fund alternatives with the following restrictions. For diversification, no more than $25,000 can be invested in any one fund. If a fund is chosen for investment, then at least $10,000 will be invested in it. No more than two of the funds can be pure growth funds, and at least one pure bond fund must be selected. The total amount invested in pure bond funds must be at least as much as the amount invested in pure growth funds. Using the following expected returns, formulate and solve a model that will determine the investment strategy that will maximize expected annual return. What assumptions have you made in your model? How often would you expect to run your model? **LO 3, 4, 5, 7**

DATA *file*
mutual_fund_planning

Fund	Type	Expected Return (%)
1	Growth	6.70
2	Growth	7.65
3	Growth	7.55
4	Growth	7.45
5	Growth & Income	7.50
6	Growth & Income	6.45
7	Growth & Income	7.05
8	Stock & Bond	6.90
9	Bond	5.20
10	Bond	5.90

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

23. **New Product Design.** Wegryn Consumer Goods (WCG) is considering entering the dish soap market with a new product it has yet to design. Using conjoint analysis, WCG has estimated part-worth utilities for 80 survey respondents for the following four attributes for dish soap:

- Surfactants—chemicals that help suspend the oils and residue on dishes (5 levels)
- Solubility—how quickly a soap dissolves in water (10 levels)
- Color—the color of the soap (5 levels)
- Scent—the aroma of the soap (4 levels)

WCG has also collected the profile of the status quo product for each of the respondents in the form of a binary vector of length 24. For example, the following vector indicates that this respondent is currently using a product comprised of level 1 of surfactants, level 6 of solubility, level 3 of color, and level 1 of scent.

DATA *file*
dishsoap

[1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 0]

The utility of the status quo product for each respondent can be obtained by adding up the part-worth's of the levels indicated for the status quo product.

The part-worth utilities and the status quo vector for each of the 110 respondents are given in the file *dishsoap*.

WCG wants to construct a new dish soap product so that the new product is more valuable than the status quo product for the maximum number of respondents. What is the optimal objective function value? **LO 2, 4, 5, 7**

24. **Manufacturing Plant Location (Revisited).** In this problem, we revisit the Martin-Beck plant location problem described in Problem 10. The management of Martin-Beck has decided to do a clean-sheet analysis. Rather than assume that the St. Louis plant is fixed as open, management wants to run a model that allows for any plant or set of plants to be open so that total cost is minimized. The annual fixed cost and the capacities of the proposed plants have been estimated (see Problem 10). The variable costs for the proposed plant locations are estimated to be the following: Detroit ($4.26), Toledo ($4.19), Denver ($4.69), and Kansas City ($4.20).

DATA *file*
stlouis_mb

We need an estimate of the fixed cost and variable cost at the current St. Louis plant. The file *stlouis_mb* contains 15 observations from previous years that will allow us to estimate these. **LO 2, 4, 5, 7**

a. Use simple linear regression, with total cost as the dependent variable (y) and volume as the independent variable (x). The model $y = b_0 + b_1x$ will give you estimates of b_1 = the per unit variable cost and b_0 = annual fixed cost. Round the variable cost estimate to the nearest cent.
b. Using the data from Problem 10, the values of variable cost given above for the proposed locations and the fixed and variable costs for St. Louis estimated in part (a), build an optimization model to minimize total cost to meet demand. Which plants are open and what is the total cost to meet demand?

25. **Second Best Manufacturing Locations.** Refer to Problem 24. Find the second-best solution. How does it compare to the optimal solution found in Problem 24? Why might the second-best solution be preferred? **LO 5, 8**

26. **Promotion Planning for a Food Bank.** The Harvest Moon Food Bank is planning its promotional events for the next year. Harvest would like to maximize the number of meals it is able to provide based on these promotions. To do this, the food bank must effectively utilize the resources it has on-hand for holding promotional events. The table below shows the relevant resources required per event for the three categories of events (Church Food Drive, 5K Run, and Concert).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Resources Required per Event per Category

	Church Food Drive	5K Run	Concert	Available (per year)
Paid Staff (hours)	110	145	138	2,000
Volunteer Staff (hours)	83	56	227	1,800
Supplies ($)	500	1,500	800	100,000

Below is the number of meals provided per event for each category and the lower and upper limit on events in the category if that category is chosen.

	Church Food Drive	5K Run	Concert
Meals Yielded per event:	143,000	90,000	113,200
Minimum Number of Events	5	2	2
Maximum Number of Events	10	4	4

Each category utilized requires a category coordinator to manage the events of that category. Harvest Moon only has two category coordinators, so this limits the number of categories chosen to exactly two. Therefore, Harvest wishes to select two of the three categories and decide how many events to hold in each of the chosen categories to maximize the number of meals it can provide during the coming year. **LO 1, 4, 5, 6, 7**

a. Develop a mixed-integer linear model to tell which two categories to use and the number of events in that category to hold so as to maximize the meals provided over the year.
b. Create a spreadsheet model of your model developed in part (a) and solve it using Excel Solver.
c. Which categories are selected and how many events of each category will be held?
d. What is the solution if we add a third category manager and all other available resources stay the same? How many more meals are provided by adding a third-category coordinator?

27. **Promotion Planning for a Food Bank (Revisited).** Refer to Problem 26. The problem involves the choice of promotions for a food bank so as to maximize the number of meals provided. **LO 5, 8**
a. Use equation (15.1) with the binary variables in your model to find the second best solution to the model from Problem 26 with two category coordinators.
b. Compare the solution you obtained in Problem 26 with two category coordinators, with the second best solution found in part (a).
c. Which solution would you recommend?

28. **Carrier Selection (Revisited).** Refer to Problem 9. The problem involves the choice of carriers to ship goods for Offhaus Manufacturing. The file *offhaus* contains bids for shipping a truckload of goods from Dayton, Ohio to each of 20 cities. Also given is the demand in truckloads for each city. Recall that each city must be assigned to only one carrier, blanks in the data set indicate the carrier did not bid on that city, and every city must be assigned to a carrier. Suppose Offhaus management has estimated that every carrier will need a program manager at a fixed cost of $50,000 per carrier used. **LO 2, 4, 5, 8**

DATA *file* offhaus

a. Develop a model that will yield the set of carriers and city assignments that will minimize shipping cost plus the fixed costs of managing the carriers.
b. Build and solve a spreadsheet model of this problem.
c. How many carriers are selected and what is the total cost?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

d. Use equation (15.1) with the binary variables that select the carriers to find the second best solution.
e. Compare the solutions from parts (b) and (d). Which would you recommend?

29. **Baseball Start Times.** The Boston Red Sox major league baseball team decided to study which game starting times they should utilize in the next season (source: *Sports Business Journal,* January 24, 2022). One way to approach this scheduling problem is to conduct a large survey of people who watch television and ask if they would watch a game starting at a particular time on a given day of the week. As an example of the data, see the matrix below where each row corresponds to a survey respondent and each column corresponds to a proposed day and start time (WN = week night, Sat = Saturday, and Sun = Sunday). All starting times are p.m. A "1" indicates the person prefers to watch a game starting at that time and a "0" indicates the person does not prefer a game at that time.

DATA *file* gametimes

	WN	WN	Sat	Sat	Sat	Sun	Sun	Sun
Respondent	6:30	7:30	2:00	6:00	8:00	2:00	4:00	7:30
1	1	1	1	1	0	0	0	1
2	0	1	1	1	1	0	0	0
3	0	1	0	0	0	1	1	0
4	1	1	0	0	0	0	0	0
5	0	0	1	0	0	1	1	1

We must choose one weeknight starting time and one starting time each for Saturday and Sunday. Sponsorship revenue for the televised games is dependent on viewership measured by the number of viewers who will watch at least one game per week. The file *gametimes* contains data from 75 respondents. **LO 2, 4, 5, 7**

a. Develop a binary integer linear model to select starting times to maximize the number of viewers for whom at least one of their preferred times have been chosen.
b. Create a spreadsheet model of your model in part (a) and solve it using Excel Solver.
c. Give the optimal starting times and the optimal number of viewers who watch at least one game per week.

30. **Baseball Start Times (Revisited).** Refer to Problem 29. The problem involves the choice of starting times for televised baseball games to maximize the number of viewers who will watch at least one game. **LO 5, 7**

DATA *file* gametimes

a. Suppose we add the constraint that at least one day-time starting time (Saturday 2:00, Sunday 2:00 or Sunday 4:00) must be chosen. What is the new solution?
b. Compare the solution you obtained in Problem 29 with the solution found in part (a). Compare the objective function values, that is, number of viewers who watch at least one game. Also compare the total number of games viewed (assume four week night games).
c. Which solution would you recommend?

Case Problem 1: Applecore Children's Clothing

Applecore Children's Clothing is a retailer that sells high-end clothes for toddlers (ages 1 to 3), primarily in shopping malls. Applecore also has a successful Internet-based sales division. Recently Dave Walker, vice-president of the e-commerce division, has been given the directive to expand the company's Internet sales. Walker commissioned a major study on the effectiveness of Internet ads placed on news web sites. The results were favorable:

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Current patrons who purchased via the Internet and saw the ads on news web sites spent more, on average, than did comparable Internet customers who did not see the ads.

With this new information on Internet ads, Walker continued to investigate how new Internet customers could most effectively be reached. One of these ideas involved strategically purchasing ads on news web sites prior to and during the holiday season. To determine which news sites might be the most effective for ads, Walker conducted a follow-up study. An e-mail questionnaire was administered to a sample of 1,200 current Internet customers to ascertain which of 30 news sites they regularly visit. The idea is that web sites with high proportions of current customer visits would be viable sources of future customers for Applecore products.

Walker would like to ascertain which news sites should be selected for ads. The problem is complicated because Walker does not want to count multiple exposures. So, if a respondent visits multiple sites with Applecore ads or visits a given site multiple times, that respondent should be counted as reached but not more than once. In other words, a customer is considered reached if they have visited at least one web site with an Applecore ad.

Data from the customer e-mail survey have begun to trickle in. Walker wants to develop a prototype model based on the current survey results. So far, 53 surveys have been returned. To keep the prototype model manageable, Walker wants to proceed with model development using the data from the 53 returned surveys and using only the first 10 news sites in the questionnaire. The costs of ads per week for the 10 web sites are given in the following table, and the budget is $10,000 per week. For each of the 53 responses received, the 10 web sites visited regularly are shown below. For a given customer–web site pair, a one indicates that the customer regularly visits that web site and a zero indicates that the customer does not regularly visit that site. **LO 2, 4, 5, 7, 8**

DATA*file*
applecore

Data for Applecore Customer Visits to News Web Sites (respondents 5 to 33 hidden)

	Web Site									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Cost/Wk ($000)	$5.0	$8.0	$3.5	$5.5	$7.0	$4.5	$6.0	$5.0	$3.0	$2.2
	Web Site									
Customer	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	1
2	1	0	0	1	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0
4	0	0	0	0	1	1	0	0	0	0
34	0	0	0	1	1	0	0	0	0	0
35	1	0	0	0	1	1	0	0	0	0
36	1	0	1	0	0	0	0	0	0	0
37	0	0	1	0	1	0	0	1	0	0
38	0	0	1	0	0	0	0	0	0	0
39	0	1	0	0	0	0	1	0	0	0
40	0	1	0	0	0	0	1	0	0	0
41	0	0	0	0	0	0	1	0	0	0
42	0	0	0	1	1	1	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	1	0	0	0	0	1
45	1	1	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	1	0	0	0
47	1	0	0	0	1	0	0	0	0	1
48	0	0	1	0	0	0	0	0	0	0
49	1	0	1	1	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0
51	0	1	0	0	0	1	0	0	0	0
52	0	0	0	0	0	0	0	0	0	0
53	0	1	0	0	1	0	0	1	1	1

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Managerial Report

1. Develop a model that will allow Applecore to maximize the number of customers reached for a budget of $10,000 for one week of promotion.
2. Solve the model. What is the maximum number of customers reached for the $10,000 budget?
3. Perform a sensitivity analysis on the budget for values from $5,000 to $35,000 in increments of $5,000. Construct a graph of percentage reach versus budget. Is the additional increase in percentage reach monotonically decreasing as the budget allocation increases? Why or why not? What is your recommended budget? Explain.

Case Problem 2: Yeager National Bank

Using aggressive mail promotion with low introductory interest rates, Yeager National Bank (YNB) built a large base of credit card customers throughout the continental United States. Currently, customers who mail their payments to YNB send their regular payments to the bank's corporate office located in Charlotte, North Carolina. Daily collections from customers making their regular payments are substantial, with an average of approximately $600,000. YNB estimates that it makes about 15% on its funds, and it would like to ensure that customer payments are credited to the bank's account as soon as possible. For instance, if it takes five days for a customer's payment to be sent through the mail, processed, and credited to the bank's account, YNB has potentially lost five days' worth of interest income. Although the time needed for this collection process cannot be eliminated, reducing it can be beneficial given the large amounts of money involved.

Instead of having all its credit card customers send their payments to Charlotte, YNB is considering having customers send their payments to one or more regional collection centers, referred to in the banking industry as lockboxes. Four lockbox locations have been proposed: Phoenix, Salt Lake City, Atlanta, and Boston. To determine which lockboxes to open and where lockbox customers should send their payments, YNB divided its customer base into five geographical regions: Northwest, Southwest, Central, Northeast, and Southeast. Every customer in the same region will be instructed to send their payment to the same lockbox. The following table shows the average number of days it takes before a customer's payment is credited to the bank's account when the payment is sent from each of the regions to each of the potential lockboxes. **LO 2, 4, 5, 6, 7**

	Location of Lockbox				
Customer Zone	Phoenix	Salt Lake City	Atlanta	Boston	Daily Collection ($1,000s)
Northwest	4	2	4	4	80
Southwest	2	3	4	6	90
Central	5	3	3	4	150
Northeast	5	4	3	2	180
Southeast	4	6	2	3	100

Managerial Report

Terry Wolff, the vice president for cash management, asked you to prepare a report containing your recommendations for the number of lockboxes and the best lockbox locations. Terry is primarily concerned with minimizing lost interest income, but wants you to also consider the effect of an annual fee charged for maintaining a lockbox at any location. Although the amount of the fee is unknown at this time, we can assume that the fees will be in the range of $20,000 to $30,000 per location. Once good potential locations have been selected, Terry will inquire as to the annual fees.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 16

Nonlinear Optimization Models

Contents

Analytics In Action:
Intercontinental Hotels

16.1 A Production Application: Par, Inc. Revisited
An Unconstrained Problem
A Constrained Problem
Solving Nonlinear Optimization Models Using Excel Solver
Sensitivity Analysis and Shadow Prices in Nonlinear Models

16.2 Local and Global Optima
Overcoming Local Optima with Excel Solver

16.3 A Location Problem

16.4 Markowitz Portfolio Model

16.5 Adoption of a New Product: The Bass Forecasting Model

16.6 Heuristic Optimization Using Excel's Evolutionary Method

Summary 877
Glossary 877
Problems 878

Available in the Cengage eBook:
Appendix: Nonlinear Programming with R

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Formulate nonlinear optimization problems by writing out the objective function and the applicable constraints.

LO 2 Construct a spreadsheet model of nonlinear optimization problems, use Excel Solver with the GRG Nonlinear method to solve the model, and interpret the results.

LO 3 Use the settings in Excel Solver to increase the likelihood of finding a global optimal solution for nonlinear optimization models.

LO 4 Formulate financial return measures and financial risk measures such as variance and semivariance in a nonlinear optimization model.

LO 5 Use nonlinear optimization models to create an efficient frontier.

LO 6 Perform sensitivity analysis for nonlinear optimization problems.

LO 7 Construct a spreadsheet model of a nonlinear optimization problem with non-smooth functions, use Excel Solver with the Evolutionary method to solve the model, and interpret the results.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

AESA*

Agencia Estatal de Seguridad Aerea (AESA) is the national aviation safety and security agency in Spain. AESA is charged with improving aviation safety in Spain, and it oversees around 50 airports, 44 airlines, and 40,000 aviation-related personnel. Each year AESA must deal with about 30,000 safety occurrences, which are divided into 88 different types including ground collisions, runway incursions, and bird strikes to aircraft. Approximately 40 of the 30,000 safety occurrences each year are classified as accidents that result in aircraft destruction or deaths.

AESA partnered with the Spanish Royal Academy of Sciences to develop a nonlinear optimization model to allocate inspection resources with the goal of reducing the number of safety occurrences, their severities, and the resulting consequences. The model assigns safety resources (workers and equipment to perform inspections) to predicted safety occurrences. The predicted safety occurrences are based on models that forecast the number of occurrences and their expected loss. These forecasts become inputs into the nonlinear optimization model. The objective of the nonlinear optimization model is to minimize the expected costs due to the forecasted safety occurrences. Because many occurrences have minimum inspection requirements, these form constraints in the model. The nonlinear optimization model is solved using R, an open-source, statistical programming language. The model output determines the safety resource allocation plan for the next year.

It is estimated that the use of the nonlinear optimization model has reduced the number of aviation accident deaths in Spain by several dozen and the costs associated with safety occurrences by more than 900 million euros per year. Furthermore, the use of standardized management science models provide more transparency and better documentation of the decisions made by the AESA.

*Based on Verónica Elvira, Francisco Bernal, Pablo Hernandez-Coronado, Esperanza Herraiz, Cesar Alfaro, Javier Gomez, and David Rios Insua, "Safer Skies over Spain," *INFORMS Journal on Applied Analytics* 50, no. 1 (2020): 21–36.

Many business processes behave in a nonlinear manner. For example, the price of a bond is a nonlinear function of interest rates, and the price of a stock option is a nonlinear function of the price of the underlying stock. The marginal cost of production often decreases with the quantity produced, and the quantity demanded for a product is usually a nonlinear function of the price. These and many other nonlinear relationships are present in many business applications.

A **nonlinear optimization problem** is any optimization problem in which at least one term in the objective function or a constraint is nonlinear. In this chapter, we examine a production problem in which the objective function is a nonlinear function of the decision variables. We discuss issues that make nonlinear optimization very different from linear optimization. We present a nonlinear model for facility location as well as the Nobel Prize–winning Markowitz model for managing the trade-off between risk and return in the construction of an investment portfolio. We also consider a well-known model that effectively forecasts sales or adoptions of a new product. In the online appendix that accompanies this chapter, we introduce the use of R to solve nonlinear programs.

16.1 A Production Application: Par, Inc. Revisited

We introduce constrained and unconstrained nonlinear optimization problems by considering an extension of the Par, Inc. linear program introduced in Chapter 14. We first consider the case in which the relationship between price and quantity sold causes the objective function to be nonlinear. The resulting unconstrained nonlinear program is then solved. As we shall see, the unconstrained optimal solution does not satisfy the production constraints of the original problem. Adding the production constraints back into the problem allows us to show the formulation and solution of a constrained nonlinear optimization model.

An Unconstrained Problem

Let us consider a revision of the Par, Inc. problem discussed in Chapter 14. Recall that Par, Inc. decided to manufacture standard and deluxe golf bags. In formulating the linear

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

programming model for the Par, Inc. problem, we assumed that the company could sell all of the standard and deluxe bags it could produce. However, depending on the price of the golf bags, this assumption may not hold. An inverse relationship usually exists between price and demand. As price increases, the quantity demanded decreases. Let P_S denote the price Par, Inc. charges for each standard bag and P_D denote the price for each deluxe bag. Assume that the demand for standard bags, S, and the demand for deluxe bags, D, are given by

$$S = 2{,}250 - 15P_S \tag{16.1}$$

$$D = 1{,}500 - 5P_D \tag{16.2}$$

The revenue generated from standard bags is the price of each standard bag, P_S, times the number of standard bags sold, S. If the cost to produce a standard bag is \$70, then the cost to produce S standard bags is $70S$. Thus, the profit contribution for producing and selling S standard bags (revenue − cost) is

$$P_S S - 70S = (P_S - 70)S \tag{16.3}$$

We can solve equation (16.1) for P_S to show how the price of a standard bag is related to the number of standard bags sold: $P_S = 150 - (1/15)S$. Substituting $150 – (1/15)S$ for P_S in equation (16.3), the profit contribution for standard bags is

$$(P_S - 70)S = [150 - (1/15)S - 70]S = 80S - (1/15)S^2 \tag{16.4}$$

Suppose that the cost to produce each deluxe golf bag is \$150. Using the same logic we used to develop equation (16.4), the profit contribution for deluxe bags is

$$(P_D - 150)D = [300 - (1/5)D - 150]D = 150D - (1/5)D^2$$

Total profit contribution is the sum of the profit contribution for standard bags and the profit contribution for deluxe bags. Thus, total profit contribution is written as

$$\text{Total profit contribution} = 80S - (1/15)S^2 + 150D - (1/5)D^2 \tag{16.5}$$

Note that the two linear demand functions, equations (16.1) and (16.2), give a nonlinear total profit contribution function, equation (16.5). This function is an example of a **quadratic function** because the nonlinear terms have an exponent of 2 (S^2 and D^2).

Details of how to use Excel Solver for nonlinear optimization are discussed in the next section.

Using Excel Solver, we find that the values of S and D that maximize the profit contribution function are $S = 600$ and $D = 375$. The corresponding prices are \$110 for standard bags and \$225 for deluxe bags, and the profit contribution is \$52,125. If all production constraints are also satisfied, these values provide the optimal solution for Par, Inc.

A Constrained Problem

In calculating the unconstrained optimal solution, we have ignored the production constraints discussed in Chapter 14. Recall that Par, Inc. has limited amounts of time available in each of four departments (cutting and dyeing, sewing, finishing, and inspection and packaging). We must enforce constraints that ensure that the amount of time used does not exceed the amount of time available in each of these departments. The problem that Par, Inc. must solve is to maximize the total profit contribution subject to all of the departmental labor hour constraints given in Chapter 14. The complete mathematical model for the Par, Inc. constrained nonlinear maximization problem is as follows:

$$\text{Max} \quad 80S - \tfrac{1}{15}S^2 + 150D - \tfrac{1}{5}D^2$$

s.t.

$$\begin{aligned}
\tfrac{7}{10}S + 1D &\le 630 && \text{Cutting and dyeing}\\
\tfrac{1}{2}S + \tfrac{5}{6}D &\le 600 && \text{Sewing}\\
1S + \tfrac{2}{3}D &\le 708 && \text{Finishing}\\
\tfrac{1}{10}S + \tfrac{1}{4}D &\le 135 && \text{Inspection and packaging}\\
S, D &\ge 0
\end{aligned}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

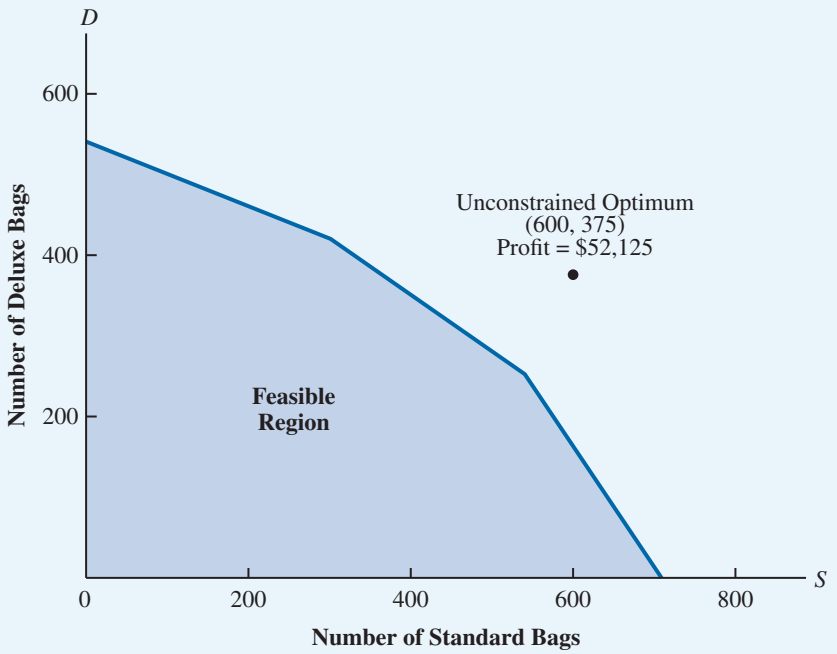
Figure 16.1 shows the feasible region for the original Par, Inc. problem, along with the unconstrained optimal solution point (600, 375). The unconstrained optimum of (600, 375) is obviously outside the feasible region.

This maximization problem is exactly the same as the Par, Inc. problem in Chapter 14 except for the nonlinear objective function. The solution to this new constrained nonlinear maximization problem is shown in Figure 16.2.

In Figure 16.2 we see three profit contribution contour lines. Each point on the same contour line is a point of equal profit. Here, the contour lines show profit contributions of $45,000, $49,920.55, and $51,500. In the original Par, Inc. problem described in Chapter 14, the objective function is linear, and thus the profit contours are straight lines. However, for the Par, Inc. problem with a quadratic objective function, the profit contours are ellipses.

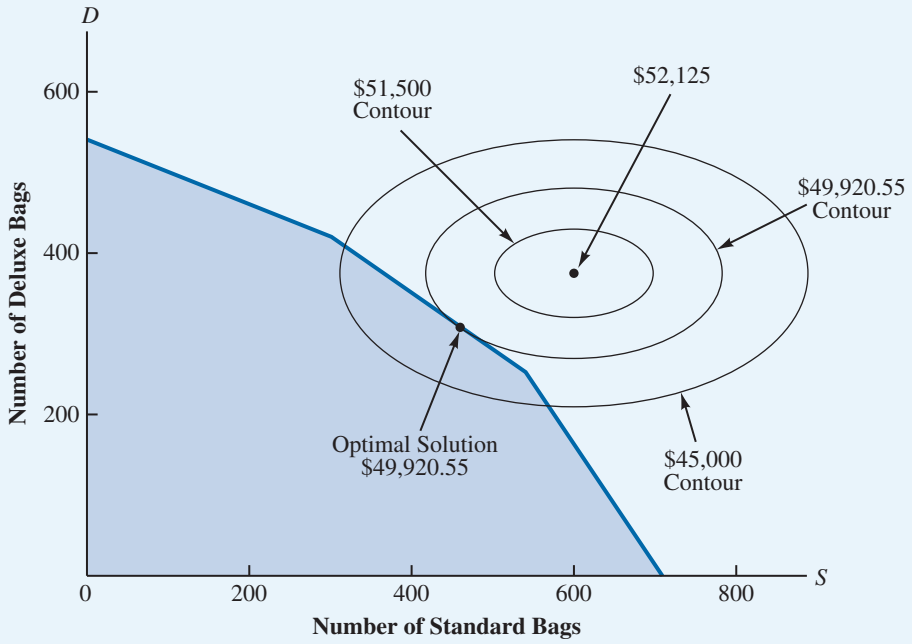
Because part of the $45,000 profit contour line cuts through the feasible region, we know that an infinite number of combinations of standard and deluxe bags will yield a profit of $45,000. An infinite number of combinations of standard and deluxe bags also provide a profit of $51,500. However, none of the points on the $51,500 contour profit line is in the feasible region. As the contour lines move farther out from the unconstrained optimum of (600, 375) the profit contribution associated with each contour line decreases. The contour line representing a profit of $49,920.55 intersects the feasible region at a single point. Without showing all of the details in solving for this point, the point of intersection is 459.717 standard bags and 308.198 deluxe bags. This solution provides the maximum possible profit. No contour line that has a profit contribution greater than $49,920.55 will intersect the feasible region. Because the contour lines are nonlinear, the contour line with the highest profit can touch the boundary of the feasible region at any point, not just an extreme point. In the Par, Inc. case, the optimal solution is on the cutting and dyeing constraint line partway between two extreme points.

Figure 16.1 The Par, Inc. Feasible Region and the Optimal Solution for the Unconstrained Optimization Problem



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 16.2 The Par, Inc. Feasible Region with Objective Function Contour Lines



It is also possible for the optimal solution to a nonlinear optimization problem to lie in the interior of the feasible region. For instance, if the right-hand sides of the constraints in the Par, Inc. problem were all increased by a sufficient amount, the feasible region would expand so that the optimal unconstrained solution point of (600, 375) with a profit contribution of $52,125 in Figure 16.2 would be in the interior of the feasible region.

Many linear optimization algorithms (e.g., the simplex method) optimize by examining only the extreme points and selecting the extreme point that gives the best solution value. As the solution to the constrained nonlinear problem for Par, Inc. illustrates, such a method will not work in the nonlinear case because the optimal solution is generally not an extreme-point solution. Hence, nonlinear optimization algorithms are more complex than linear optimization algorithms, and the details are beyond the scope of this text. Fortunately, we do not need to know how nonlinear algorithms work; we just need to know how to use them. Computer software such as Excel Solver and Analytic Solver are available to solve nonlinear optimization problems.

Next we discuss how to use Excel Solver to solve nonlinear optimization problems.

Solving Nonlinear Optimization Models Using Excel Solver

We use the constrained nonlinear problem for Par, Inc. to illustrate how to use Excel Solver to solve nonlinear optimization problems. The procedure for developing and entering the model in Excel is the same as for linear problems as discussed in Chapter 14, except that one or more of the functions is nonlinear.

Figure 16.3 shows the Excel model and Solver dialog box for the nonlinear Par, Inc. problem. The SUMPRODUCT function is used in cells B19 through B22 to calculate the number of hours required in each department. The price function for standard bags is entered in cell

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

B25 as *=150-(1/15)*$B$14* and similarly for deluxe bags in cell D26 as *=300-(1/5)*$C$14*. The objective function in cell B16 contains the formula *=(B25-B9)*B14+(B26-C9)*C14*, which corresponds to $(150 - (1/15)S - 70)S + (300 - (1/5)D - 150)D$. As previously shown, this is mathematically equivalent to equation (16.5) because $(150 - (1/15)S - 70)S + (300 - (1/5)D - 150)D = 80S - (1/15)S^2 + 150D - (1/5)D^2$.

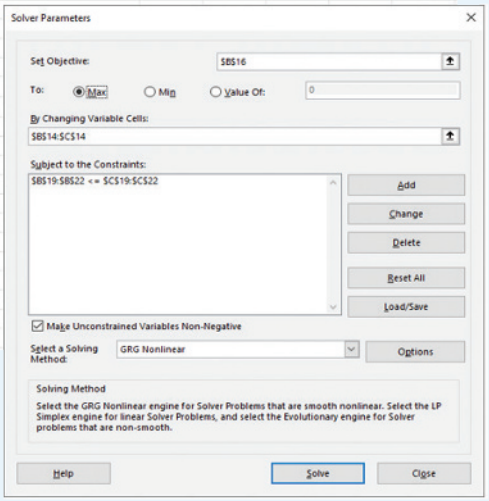
To invoke Solver, we follow these steps:

Step 1. Click the **Data** tab in the Ribbon
Step 2. Click **Solver** in the **Analyze** group
Step 3. When the **Solver Parameters** dialog box appears:
Enter *B16* into the **Set Objective:** box
Step 4. Enter *B14:C14* into the **By Changing Variable Cells:** box area
Step 5. Click the **Add** button
Enter *B19:B22* in the **Cell Reference:** box
Select ≤ from the drop-down menu
Enter *C19:C22* in the **Constraint:** box
Click **OK**
Step 6. Select the **Make Unconstrained Variables Non-negative** option
Step 7. For **Select a Solving Method:** select **GRG Nonlinear** from the drop-down menu
Step 8. Click **Solve**
Step 9. When the **Solver Results** dialog box appears, click **OK**

Figure 16.3 Spreadsheet Model and Solver Parameters Dialog Box for the Nonlinear Par, Inc. Problem

MODEL file
parnonlinear

	A	B	C	D
1	**Par, Inc.**			
2	**Parameters**			
3		**Production Time (Hours)**		**Time Available**
4	**Operation**	**Standard**	**Deluxe**	**(Hours)**
5	Cutting and Dyeing	0.7	1	630
6	Sewing	0.5	0.833333333	600
7	Finishing	1	0.666666667	708
8	Inspection and Packaging	0.1	0.25	135
9	Marginal Cost	$70.00	$150.00	
10				
11	**Model**			
12				
13		**Standard**	**Deluxe**	
14	**Bags Produced**	459.717	308.198	
15				
16	**Total Profit**	$49,920.55		
17				
18	**Operation**	**Hours Used**	**Hours Available**	
19	Cutting and Dyeing	630.00	630	
20	Sewing	486.69	600	
21	Finishing	665.18	708	
22	Inspection and Packaging	123.02	135	
23				
24				
25	Standard Bag Price Function	119.35		
26	Deluxe Bag Price Function	238.36		



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The complete model for the constrained nonlinear Par, Inc. problem is contained in the file *parnonlinear*.

Note that Par, Inc., cannot produce fractional bags, but we can round this solution to 460 Standard bags and 308 Deluxe bags. This rounded solution is feasible and results in an objective function value of $49,920.23.

The Answer Report generated by Excel Solver has the same structure as that of linear programs. Rather than show the Answer Report here, we refer to the optimal values shown in the spreadsheet in Figure 16.3. The optimal value of the objective function is $49,920.55, and this is achieved by producing 459.717 standard bags and 308.198 deluxe bags. This is the optimal point shown geometrically in Figure 16.2. Comparing cells C19 through C22 with cells D19 through D22 shows that only the cutting and dyeing constraint is binding, which is consistent with Figure 16.2. The optimal prices, based on the optimal quantities, are shown in cells B25 and B26. The optimal price for a standard bag is $119.35 and the optimal price for a deluxe bag is $238.36.

Sensitivity Analysis and Shadow Prices in Nonlinear Models

The Sensitivity Report for the nonlinear Par, Inc. problem is shown in Figure 16.4. As in the linear case, there are two sections: one for the variables and the other for constraints. The variables section gives the cell location, name, final (optimal) value, and **reduced gradient** for each variable. The reduced gradient is analogous to the reduced cost for linear models. It is essentially the shadow price of the nonnegativity constraint or, more generally, the shadow price of a binding simple lower or upper bound on the decision variable.

The constraint section gives the cell location, name, and final value for the left-hand side of each constraint. For the Par, Inc. problem, the final values are the amount of time in hours used in each of the four departments. The far right column gives the **Lagrangian multiplier** for each constraint. The Lagrangian multiplier is the shadow price for a constraint in a nonlinear problem. In other words, the Lagrangian multiplier is the rate of change of the objective function with respect to the right-hand side of a constraint.

Figure 16.4 Excel Solver Sensitivity Report for the Nonlinear Par, Inc. Problem

Variable Cells

Cell	Name	Final Value	Reduced Gradient
$B$14	Bags Produced Standard	459.7166	0
$C$14	Bags Produced Deluxe	308.1984	0

Constraints

Cell	Name	Final Value	Lagrange Multiplier
$B$19	Cutting and Dyeing Hours Used	630.0000	26.7206
$B$20	Sewing Hours Used	486.6903	0
$B$21	Finishing Hours Used	665.1822	0
$B$22	Inspection and Packaging Hours Used	123.0213	0

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For the Par, Inc. example, as we increase the number of hours available in the cutting and dyeing department, we expect the profit to increase by $26.72 per hour. However, notice that no ranges are given for allowable changes to the right-hand side. This is because the allowable increase and decrease are essentially zero. Changing the right-hand side of a binding constraint by even a small amount will change the value of Lagrangian multiplier. Nonetheless, the Lagrangian multiplier does give an estimate of the importance of relieving a binding constraint.

16.2 Local and Global Optima

The neighborhood of a solution is a mathematical concept that refers to the set of points within a relatively close proximity of the solution. See Figure 16.7 for a graphical example of local minimums and local maximums.

A feasible solution is a **local optimum** if no other feasible solution with a better objective function value is found in the immediate neighborhood. For example, for the constrained Par, Inc. problem, the local optimum corresponds to a local maximum; a point is a **local maximum** if no other feasible solution with a larger objective function value is in the immediate neighborhood. Similarly, for a minimization problem, a point is a **local minimum** if no other feasible solution with a smaller objective function value is in the immediate neighborhood.

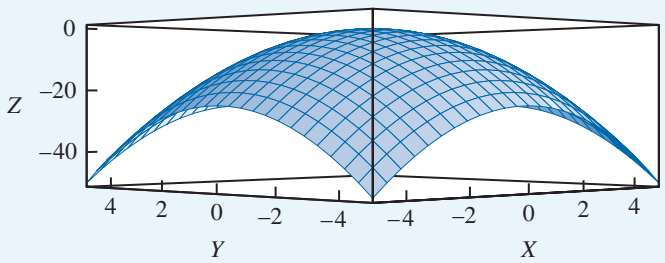
Nonlinear optimization problems can have multiple local optimal solutions, which means we are concerned with finding the best of the local optimal solutions. A feasible solution is a **global optimum** if no other feasible point with a better objective function value exists in the feasible region. In the case of a maximization problem, the global optimum corresponds to a **global maximum**. A point is a global maximum if no other point in the feasible region gives a strictly larger objective function value. For a minimization problem, a point is a **global minimum** if no other feasible point with a strictly smaller objective function value is in the feasible region. A global maximum is also a local maximum, and a global minimum is also a local minimum.

All global optimal solutions are local optimal solutions, but not all local optimal solutions are global optimal solutions.

Nonlinear problems with multiple local optima are difficult to solve. But in many nonlinear applications, a single local optimal solution is also the global optimal solution. For such problems, we need to find only a local optimal solution. We will now present some of the more common classes of nonlinear problems of this type.

Consider the function $f(X,Y) = -X^2 - Y^2$. The shape of this function is illustrated in Figure 16.5. A function that is bowl-shaped down is called a **concave function**. The maximum value for this particular function is 0, and the point (0, 0) gives the optimal value of 0. The point (0, 0) is a local maximum; but it is also a global maximum because no point gives a larger function value. In other words, no values of X and Y result in an objective function value greater than 0. Functions that are concave, such as $f(X,Y) = -X^2 - Y^2$, have a single local maximum that is also a global maximum. This type of nonlinear problem is relatively easy to maximize.

Figure 16.5 A Concave Function $f(X,Y) = -X^2 - Y^2$



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The objective function for the nonlinear Par, Inc. problem is an example of a concave function:

$$80S - \tfrac{1}{15}S^2 + 150D - \tfrac{1}{5}D^2$$

In general, if all the squared terms in a quadratic function have a negative coefficient and there are no cross-product terms, such as xy (or for the Par, Inc. problem, SD), then the function is a concave quadratic function. Thus, for the Par, Inc. problem, we are assured that the local maximum identified by Excel Solver in Figure 16.3 is the global maximum.

Let us now consider another type of function with a single local optimum that is also a global optimum. Consider the function $f(X,Y) = X^2 + Y^2$. The shape of this function is illustrated in Figure 16.6. It is bowl-shaped up and called a **convex function**. The minimum value for this particular function is 0, and the point (0, 0) gives the minimum value of 0. The point (0, 0) is a local minimum and a global minimum because no values of X and Y give an objective function value less than 0. Convex functions, such as $f(X,Y) = X^2 + Y^2$, have a single local minimum and are relatively easy to minimize.

For a concave function, we can be assured that if our computer software finds a local maximum, it has found a global maximum. Similarly, for a convex function, we know that if our computer software finds a local minimum, it has found a global minimum. However, some nonlinear functions have multiple local optima. For example, Figure 16.7 shows the graph of the following function over the feasible region: $0 \le X \le 1, 0 \le Y \le 1$:

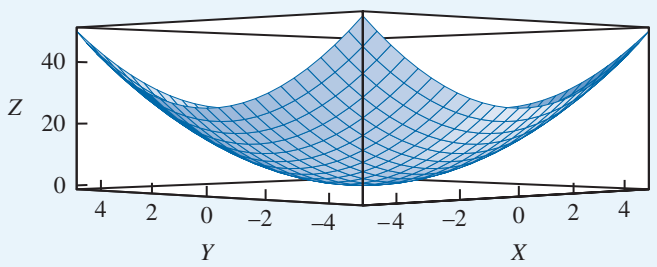
$$f(X,Y) = X\sin(5\pi X) + Y\sin(5\pi Y)$$

where sin is the trigonometric sine function, and π is approximately 3.1416. The hills and valleys in this graph show that this function has a number of local maximums and local minimums.

From a technical standpoint, functions with multiple local optima pose a serious challenge for optimization software; most nonlinear optimization software methods can get stuck and terminate at a local optimum. Unfortunately, many applications can be nonlinear with multiple local optima, and the objective function value for a local optimum may be much worse than the objective function value for a global optimum. Developing algorithms capable of finding the global optimum is currently an active area of research.

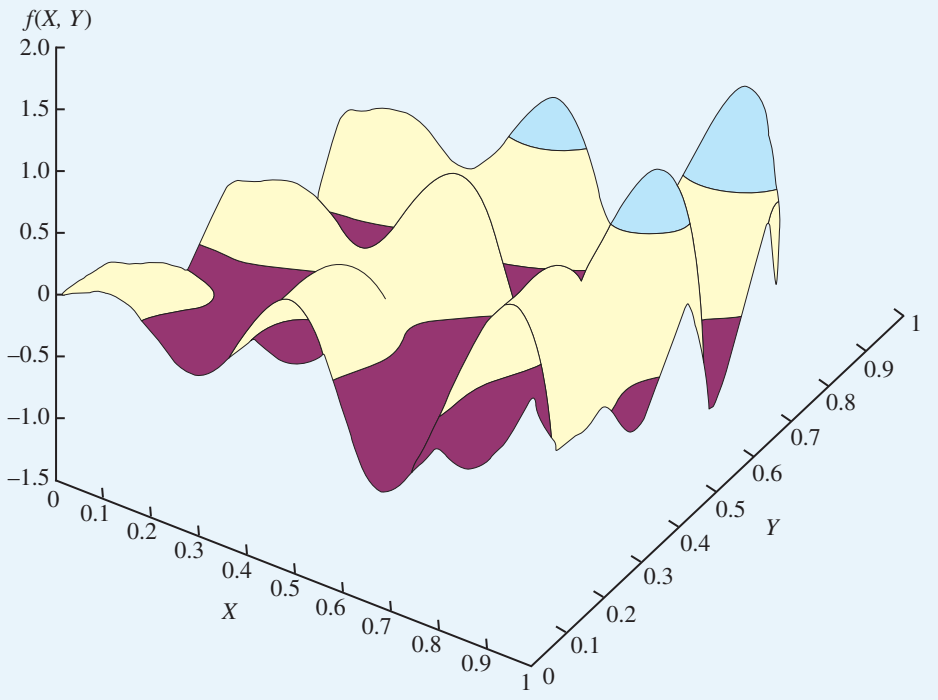
Next we discuss a very practical approach to dealing with local maximums and local minimums when using Excel Solver for nonlinear problems.

Figure 16.6 A Convex Function $f(X,Y) = X^2 + Y^2$



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 16.7 A Function with Local Maxima and Minima



Overcoming Local Optima with Excel Solver

How do you know when multiple local optima exist? The mathematical ways to determine this are beyond the scope of this text. From a practical point of view, if the solution obtained by optimization software depends on the starting point, then there are multiple local optima. Thus, when using Excel Solver, if the solution returned from Solver is different when starting from different values in the decision variable cells, then there are local optima. The converse is not necessarily true; that is, if the same solution is returned when starting from a different set of starting points, this does not necessarily mean that you have found the global optimal solution.

MODEL file
localoptima

Let us consider the problem shown in Figure 16.7:

$$\text{Max } f(X, Y) = X \sin(5\pi X) + Y \sin(5\pi Y)$$

s.t

$$0 \le X \le 1$$
$$0 \le Y \le 1$$

Table 16.1 shows the results returned from Excel Solver for different starting points (values in the decision variable cells when Solver is invoked). In each of the five cases in Table 16.1, Solver returns with the message, "Solver has converged to the current solution. All constraints are satisfied."

Excel Solver does provide an option that allows you to increase the confidence that you have found a global optimal solution. Clicking **Options** on the **Solver Parameters** dialog box and then selecting the **GRG Nonlinear** tab result in the dialog box shown in Figure 16.8. Clicking the **Use Multistart** option in the **Multistart** section causes Solver to use multiple starting solutions and report the best solution found from all of the

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 16.1 Solutions from Excel Solver for a Problem with Multiple Local Optima

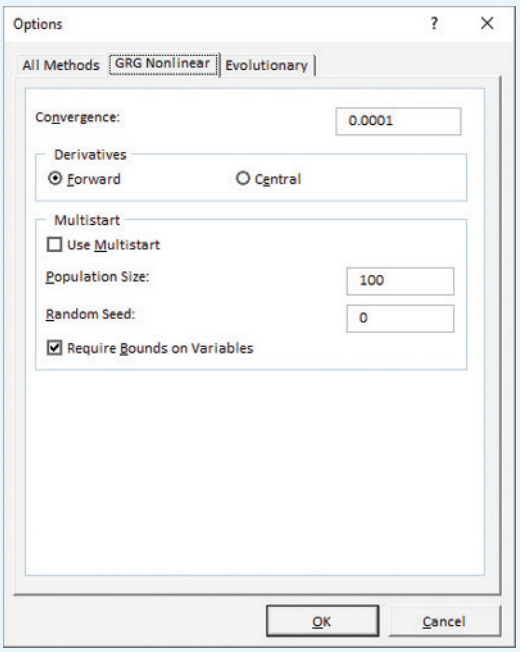
Starting Point		Solution Returned		
X	*Y*	*X*	*Y*	Objective Function Value
0.000	0.000	0.129	0.129	0.231
1.000	0.000	0.905	0.000	0.902
0.000	1.000	0.000	0.905	0.902
0.500	0.500	0.508	0.508	1.008
1.000	1.000	0.905	0.905	1.805

If the solution to a problem appears to depend on the starting values for the decision variables, we recommend you use the Multistart option.

starting points. The **Population Size** is the number of starting points used. Solver selects starting points randomly using the **Random Seed** (an integer value) such that the points are within the bounds specified. Although providing simple lower and upper bounds is not required (unless the **Require Bounds on the Variables** option is selected), the procedure is much more effective when bounds are provided. We recommend selecting the **Require Bounds on the Variables** checkbox and providing bounds before you use the Multistart option.

In Figure 16.8, randomly generated starting points will be used and simple bounds of 0 and 1 have been specified as constraints in the Solver dialog box. The result reported by Solver is $X = 0.90447$, $Y = 0.90447$, with objective function $= 1.804$. The message provided by Solver is "Solver converged in probability to a global solution."

Figure 16.8 The GRG Nonlinear Tab in Solver Options



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notes + Comments

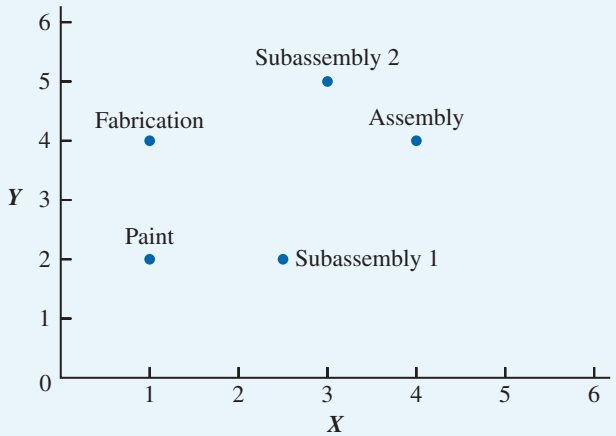
1. The Multistart option works best with bounds specified on each decision variable. It is often easy to calculate effective upper and lower bounds for the decision variables. For example, if you have a linear less-than-or-equal-to constraint with positive coefficients, upper bounds can be calculated by simply dividing the right-hand side by the coefficient for each variable. Using the cutting and dyeing constraint from the Par, Inc. problem, $\frac{7}{10}S + 1D \leq 630$, we can deduce the following upper bounds: $S \leq 630/(7/10) = 900$ and $D \leq 630/1 = 630$.
2. In addition to GRG Nonlinear, Excel Solver provides another solution method, the Evolutionary method, to solve nonlinear problems with local optimal solutions. Evolutionary Solver is based on a method that searches for an optimal solution by iteratively adjusting a population of candidate solutions. The Evolutionary method is discussed in more detail in section 16.6.

16.3 A Location Problem

Let us consider the case of LaRosa Machine Shop (LMS). LMS is studying where to locate its tool bin facility on the shop floor. The locations of the five production stations appear in Figure 16.9. In an attempt to be fair to the workers in each of the production stations, management has decided to try to find the position of the tool bin that would *minimize the sum of the distances* from the tool bin to the five production stations. We define the following decision variables:

X = horizontal location of the tool bin
Y = vertical location of the tool bin

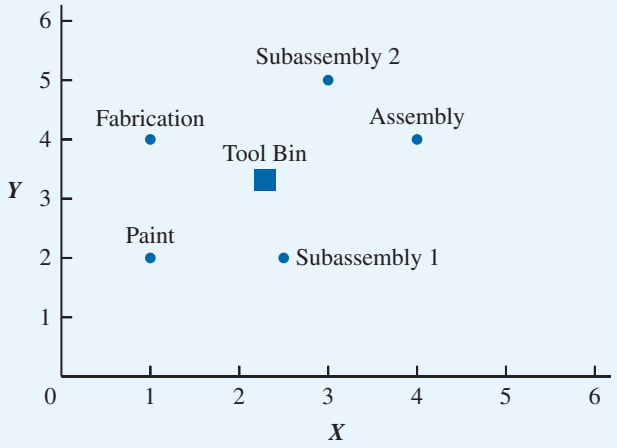
Figure 16.9 Data for the LMS Tool Bin Location Problem



Station	Location X	Location Y
Fabrication	1	4
Paint	1	2
Subassembly 1	2.5	2
Subassembly 2	3	5
Assembly	4	4

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 16.10 Solution to the LMS Tool Bin Location Problem



Station	Location X	Location Y
Fabrication	1	4
Paint	1	2
Subassembly 1	2.5	2
Subassembly 2	3	5
Assembly	4	4

We may measure the distance from a station to the tool bin located at (X, Y) by using Euclidean (straight-line) distance. For example, the distance from fabrication located at the coordinates (1, 4) to the tool bin located at the coordinates (X, Y) is given by

$$\sqrt{(X-1)^2+(Y-4)^2}$$

The unconstrained optimization problem is as follows:

MODEL file
larosa

$$\text{Min} \quad \left(\sqrt{(X-1)^2+(Y-4)^2}+\sqrt{(X-1)^2+(Y-2)^2}+\sqrt{(X-2.5)^2+(Y-2)^2}\right.$$
$$\left.+\sqrt{(X-3)^2+(Y-5)^2}+\sqrt{(X-4)^2+(Y-4)^2}\right)$$

Note that we do not require that the variables X or Y be nonnegative. The optimal solution found by Excel Solver is $X = 2.230$, $Y = 3.349$. The solution is shown in Figure 16.10.

Location models are used extensively for determining the optimal locations for everything from drilling holes in computer circuit boards to locating distribution centers and retail stores in supply chains. A variety of location models can be created by using different objective functions or by adding additional constraints on distances traveled.

16.4 Markowitz Portfolio Model

Harry Markowitz received the 1990 Nobel Prize for his ground-breaking work in portfolio optimization. The Markowitz mean-variance portfolio model is a classic application of nonlinear programming. In this section, we present the **Markowitz mean-variance portfolio model**. Money management firms throughout the world use numerous variations of this basic model.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A key trade-off in financial planning is that between risk and return. For a chance to earn greater returns, the investor must also accept greater risk. In most portfolio optimization models, the *return* used is the expected (or average) return of the possible outcomes, and the *risk* is some measure of variability in these possible outcomes. To illustrate the Markowitz portfolio model, let us consider the case of Hauck Investment Services.

Hauck Investment Services designs annuities, IRAs, 401(k) plans, and other investment vehicles for investors with a variety of risk tolerances. Hauck would like to develop a portfolio model that can be used to determine an optimal portfolio involving a mix of six mutual funds. Table 16.2 shows the annual return (%) for five 1-year periods for the six mutual funds. Year 1 represents a year in which all mutual funds yield good returns. Year 2 is also a good year for most of the mutual funds. But year 3 is a bad year for the small-cap value fund, year 4 is a bad year for the intermediate-term bond fund, and year 5 is a bad year for four of the six mutual funds.

It is not possible to predict the exact returns for any of the funds over the next 12 months, but the portfolio managers at Hauck Financial Services think that the returns for the five years shown in Table 16.2 are scenarios that can be used to represent the possibilities for the next year. For the purpose of building portfolios for their clients, Hauck's portfolio managers will choose a mix of these six mutual funds and assume that one of the five possible scenarios will describe the return over the next 12 months.

The portfolio construction problem is to determine how much of the portfolio to invest in each investment alternative. To determine the proportion of the portfolio that will be invested in each of the mutual funds we use the following decision variables:

FS = proportion of portfolio invested in the foreign stock mutual fund
IB = proportion of portfolio invested in the intermediate-term bond fund
LG = proportion of portfolio invested in the large-cap growth fund
LV = proportion of portfolio invested in the large-cap value fund
SG = proportion of portfolio invested in the small-cap growth fund
SV = proportion of portfolio invested in the small-cap value fund

Because the sum of these proportions must equal one, we need the following constraint:

$$FS + IB + LG + LV + SG + SV = 1$$

The other constraints are concerned with the return that the portfolio will earn under each of the planning scenarios in Table 16.2.

The portfolio return over the next 12 months depends on which of the possible scenarios (years 1 through 5) in Table 16.2 occurs. Let R_1 denote the portfolio return if the scenario represented by year 1 occurs, R_2 denote the portfolio return if the scenario represented by

Table 16.2 Mutual Fund Performances in Five Selected Years (Used as Planning Scenarios for the Next 12 Months)

	Annual Return (%)				
Mutual Fund	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**
Foreign Stock	10.06	13.12	13.47	45.42	−21.93
Intermediate-Term Bond	17.64	3.25	7.51	−1.33	7.36
Large-Cap Growth	32.41	18.71	33.28	41.46	−23.26
Large-Cap Value	32.36	20.61	12.93	7.06	−5.37
Small-Cap Growth	33.44	19.40	3.85	58.68	−9.02
Small-Cap Value	24.56	25.32	−6.70	5.43	17.31

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

year 2 occurs, and so on. The portfolio returns for the five planning years (scenarios) are as follows:

Scenario 1 return:

$$R_1 = 10.06FS + 17.64IB + 32.41LG + 32.36LV + 33.44SG + 24.56SV$$

Scenario 2 return:

$$R_2 = 13.12FS + 3.25IB + 18.71LG + 20.61LV + 19.40SG + 25.32SV$$

Scenario 3 return:

$$R_3 = 13.47FS + 7.51IB + 33.28LG + 12.93LV + 3.85SG - 6.70SV$$

Scenario 4 return:

$$R_4 = 45.42FS - 1.33IB + 41.46LG + 7.06LV + 58.68SG + 5.43SV$$

Scenario 5 return:

$$R_5 = -21.93FS + 7.36IB - 23.26LG - 5.37LV - 9.02SG + 17.31SV$$

If p_s is the probability of scenario s, among n possible scenarios, then the *expected* return for the portfolio is $\overline{R}$, where

$$\overline{R} = \sum_{s=1}^{n} p_s R_s \tag{16.6}$$

If we assume that the five planning scenarios in the Hauck Financial Services model are equally likely to occur, then

$$\overline{R} = \sum_{s=1}^{5} \tfrac{1}{5} R_s = \tfrac{1}{5} \sum_{s=1}^{5} R_s$$

Other common risk measures in finance include semivariance, value at risk (VaR) and conditional value at risk (CVaR).

Measuring risk is a bit more difficult. Entire books are devoted to the topic of risk measurement. The measure of risk most often associated with the Markowitz portfolio model is the variance of the portfolio's return. If the expected return is defined by equation (16.6), the variance of the portfolio's return is

$$Var = \sum_{s=1}^{n} p_s \left(R_s - \overline{R}\right)^2 \tag{16.7}$$

For the Hauck Financial Services example, the five planning scenarios are equally likely, thus

$$Var = \sum_{s=1}^{5} \tfrac{1}{5} \left(R_s - \overline{R}\right)^2$$

The portfolio variance is the average of the sum of the squares of the deviations from the mean value under each scenario. The larger this number, the more widely dispersed the scenario returns are about the average value. If the portfolio variance were equal to zero, then every scenario return R_i would be equal, and there would be no risk.

Two basic ways to formulate the Markowitz model are (1) to minimize the variance of the portfolio subject to a constraint on the expected return of the portfolio and (2) to maximize the expected return of the portfolio subject to a constraint on variance. Consider the first case. Assume that Hauck clients would like to construct a portfolio from the six mutual funds listed in Table 16.2 that will minimize their risk as measured by the portfolio variance. However, the clients also require the expected portfolio return to be at least 10%. In our notation, the objective function is

$$\text{Min} \quad \tfrac{1}{5} \sum_{s=1}^{5} \left(R_S - \overline{R}\right)^2$$

The constraint on expected portfolio return is $\overline{R} \geq 10$. The complete Markowitz model involves 12 variables and 8 constraints (excluding the nonnegativity constraints).

$$\text{Min} \tfrac{1}{5} \sum_{s=1}^{5} \left(R_S - \overline{R}\right)^2 \tag{16.8}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

s.t.

$$10.06FS + 17.64IB + 32.41LG + 32.36LV + 33.44SG + 24.56SV = R_1 \quad \textbf{(16.9)}$$

$$13.12FS + +3.25IB + 18.71LG + 20.61LV + 19.40SG + 25.32SV = R_2 \quad \textbf{(16.10)}$$

$$13.47FS + 7.51IB + 33.28LG + 12.93LV + 3.85SG - 6.70SV = R_3 \quad \textbf{(16.11)}$$

$$45.42FS - 1.33IB + 41.46LG + 7.06LV + 58.68SG + 5.43SV = R_4 \quad \textbf{(16.12)}$$

$$-21.93FS + 7.36IB - 23.26LG - 5.37LV - 9.02SG + 17.31SV = R_5 \quad \textbf{(16.13)}$$

$$FS + IB + LG + LV + SG + SV = 1 \quad \textbf{(16.14)}$$

$$\tfrac{1}{5}\sum_{s=1}^{5} R_S = \overline{R} \quad \textbf{(16.15)}$$

$$\overline{R} \geq 10 \quad \textbf{(16.16)}$$

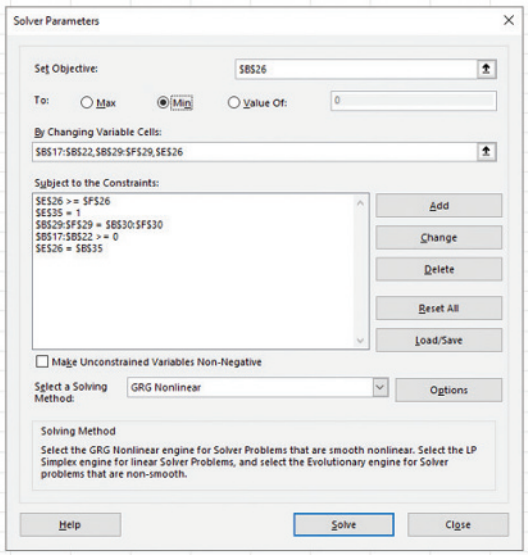
$$FS, IB, LG, LV, SG, SV \geq 0 \quad \textbf{(16.17)}$$

The objective for the Markowitz model is to minimize portfolio variance. Equations (16.9) through (16.13) define the return for each scenario. Equation (16.14) requires all of the money to be invested in the mutual funds; this constraint is often called the *unity constraint*. Equation (16.15) defines $\overline{R}$, which is the expected return of the portfolio. Equation (16.16) requires the portfolio return to be at least 10%. Finally, equation (16.17) requires a nonnegative investment in each Hauck mutual fund. Note that R_1, R_2, R_3, R_4, and R_5, as well as $\overline{R}$, are not required to be nonnegative. It is possible that the return in a given scenario or the expected return of the portfolio is negative.

The solution for this model using a required return of at least 10% appears in Figure 16.11. The minimum value for the portfolio variance is 27.136. This solution implies

Figure 16.11 Solution for the Hauck Minimum Variance Portfolio with a Required Return of At Least 10%

	A	B	C	D	E	F
1	**Hauck Financial Services - Markowitz Version**					
2	**Parameters**					
3			Planning Scenarios			
4	Mutual Fund	Year 1	Year 2	Year 3	Year 4	Year 5
5	Foreign Stock	10.06	13.12	13.47	45.42	-21.93
6	Intermediate-Term Bond	17.64	3.25	7.51	-1.33	7.36
7	Large-Cap Growth	32.41	18.71	33.28	41.46	-23.26
8	Large-Cap Value	32.36	20.61	12.93	7.06	-5.37
9	Small-Cap Growth	33.44	19.4	3.85	58.68	-9.02
10	Small-Cap Value	24.56	25.32	-6.7	5.43	17.31
11						
12	Required Return	10				
13						
14						
15	**Model**					
16						
17	Foreign Stock	0.1584				
18	Intermediate-Term Bond	0.5255				
19	Large-Cap Growth	0.0421				
20	Large-Cap Value	0.0000				
21	Small-Cap Growth	0.0000				
22	Small-Cap Value	0.2740				
23						
24						
25					Rbar	Required Return
26	Variance	27.1362			10	10
27						
28		Year 1	Year 2	Year 3	Year 4	Year 5
29	R	18.957	11.512	5.644	9.728	4.159
30	Scenario Return	18.957	11.512	5.644	9.728	4.159
31	Deviation From Mean	8.957	1.512	-4.356	-0.272	-5.841
32	Deviation From Mean Squared	80.228	2.286	18.976	0.074	34.117
33						
34					Proportion Sum	
35	Portfolio Expected Return	10			1	



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

that the clients will get an expected return of 10% $(\overline{R} \geq 10)$ and minimize their risk as measured by portfolio variance by investing approximately 16% of the portfolio in the foreign stock fund $(FS = 0.158)$, 53% in the intermediate bond fund $(IB = 0.525)$, 4% in the large-cap growth fund $(LG = 0.042)$, and 27% in the small-cap value fund $(SV = 0.274)$.

MODEL file
hauckmarkowitz

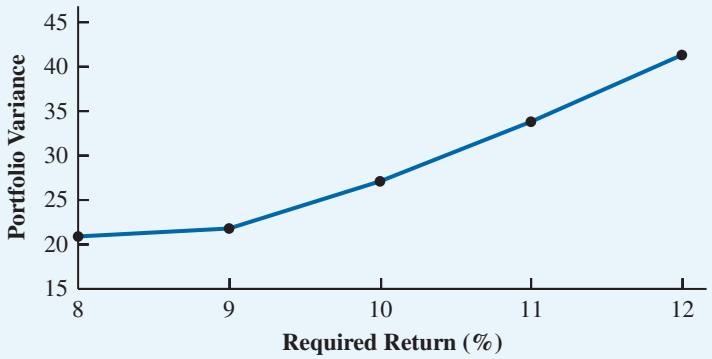
The **Solver Parameters** dialog box is also shown in Figure 16.11. Note that we have selected **GRG Nonlinear** as the method and we have *not* selected **Make Unconstrained Variables Non-Negative**. Instead we have entered as an explicit constraint set that B17 through B22 must be ≥ 0.

The Markowitz portfolio model provides a convenient way for an investor to trade off risk versus return. In practice, this model is typically solved iteratively for different values of required expected return. Figure 16.12 is a graph of the minimum portfolio variances versus required expected returns as required expected return is varied from 8% to 12% in increments of 1%. In finance, this graph is called the **efficient frontier**. Each point on the efficient frontier is the minimum possible risk (measured by portfolio variance) for the given return. By looking at the graph of the efficient frontier, investors can select the mean-variance combination with which they are most comfortable.

Notes + Comments

1. Notice that the solution given in Figure 16.11 has more than 50% of the portfolio invested in the intermediate-term bond fund. It may be unwise to allow one asset to contribute so heavily to the portfolio. Upper and lower bounds on the amount of an asset type in the portfolio can be easily modeled. Hence, upper bounds are often placed on the percentage of the portfolio invested in a single asset. Likewise, it might be undesirable to include an extremely small quantity of an asset in the portfolio. Thus, there may be constraints that require nonzero amounts of an asset to be at least a minimum percentage of the portfolio.
2. In the Hauck example, 100% of the available portfolio was invested in mutual funds. However, risk-averse investors often prefer to have some of their money in a so-called risk-free asset, such as U.S. Treasury Bills. Thus, many portfolio optimization models allow funds to be invested in a risk-free asset.
3. In this section, portfolio variance was used to measure risk. However, variance, as it is defined, counts deviations both above and below the mean. Most investors are happy with returns above the mean but wish to avoid returns below the mean. Hence, numerous portfolio models allow for flexible risk measures.
4. In practice, both brokers and mutual fund companies adjust portfolios as new information becomes available. However, constantly adjusting a portfolio may lead to large transaction costs. Extensions of the basic Markowitz model take into account transaction costs.

Figure 16.12 An Efficient Frontier for the Markowitz Portfolio Model



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

16.5 Adoption of a New Product: The Bass Forecasting Model

Forecasting new adoptions after a product introduction is an important problem in marketing. In this section, we introduce a forecasting model developed by Frank Bass[1] that has proven to be particularly effective in forecasting the adoption of innovative and new technologies in the marketplace. Nonlinear optimization is used to estimate the parameters of the Bass forecasting model. The model has three parameters that must be estimated.

m = the number of people estimated to eventually adopt the new product

A company introducing a new product is obviously interested in the value of parameter m.

q = the coefficient of imitation

Parameter q measures the likelihood of adoption due to a potential adopter being influenced by someone who has already adopted the product. It measures the word-of-mouth or social media effect influencing purchases.

p = the coefficient of innovation

Parameter p measures the likelihood of adoption, assuming no influence from someone who has already purchased (adopted) the product. It is the likelihood of someone adopting the product because of their own interest in the innovation.

Using these parameters, let us now develop the forecasting model. Let S_t be the number of new adopters in period t. Let C_{t-1} denote the number of people who have adopted the product through time $t - 1$. That is, C_{t-1} is the cumulative number of adopters through period $t - 1$ ($C_{(t-1)} = \sum_{j=1}^{t-1} s_j$). Because m is the number of people estimated to eventually adopt the product, $m - C_{t-1}$ is the number of potential adopters remaining at time $t - 1$. We refer to the time interval between time $t - 1$ and time t as period t. During period t, some percentage of the remaining number of potential adopters, $m - C_{t-1}$, will adopt the product. This value depends on the likelihood of a new adoption.

Loosely speaking, the likelihood of a new adoption is the likelihood of adoption due to imitation plus the likelihood of adoption due to innovation. The likelihood of adoption due to imitation is a function of the number of people who have already adopted the product. The larger the current pool of adopters, the greater their influence through word of mouth. Because C_{t-1}/m is the fraction of the number of people estimated to adopt the product by time $t - 1$, the likelihood of adoption due to imitation is computed by multiplying this fraction by q, the coefficient of imitation. Thus, the likelihood of adoption due to imitation is

$$q(C_{t-1}/m)$$

The likelihood of adoption due to innovation is simply p, the coefficient of innovation. Thus, the likelihood of adoption is

$$p + q(C_{t-1}/m)$$

The Bass forecasting model given in equation (16.18) can be rigorously derived from statistical principles. Rather than providing such a derivation, we have emphasized the intuitive aspects of the model.

Using the likelihood of adoption we can develop a forecast of the remaining number of potential customers who will adopt the product during time period t. Thus, F_t, the forecast of the number of new adopters during time period t, is

$$F_t = \big(p + q[C_{t-1}/m]\big)(m - C_{t-1}) \qquad \textbf{(16.18)}$$

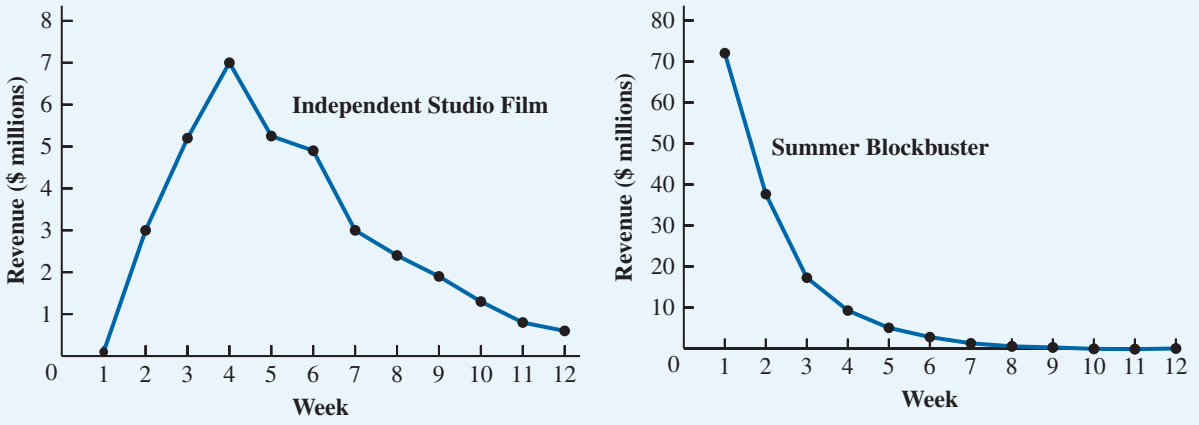
In developing a forecast of new adoptions in period t using the Bass model, the value of C_{t-1} will be known from past data. But we also need to know the values of the parameters to use in the model. Let us now see how nonlinear optimization is used to estimate the parameter values m, p, and q.

Consider Figure 16.13. This figure shows the graph of box office revenues (in $ millions) for two different films, an independent studio film and a summer blockbuster

[1]See Bass, Frank, M., "New Product Growth for Model Consumer Durables," *Management Science*, Vol. 15, No 1, pp. 215–227, 1969.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 16.13 Weekly Box Office Revenues for an Independent Studio Film and a Summer Blockbuster Movie



action movie, over the first 12 weeks after release. Strictly speaking, box office revenues for time period t are not the same as the number of adopters during time period t. However, the number of repeat customers is usually small, and box office revenues are a multiple of the number of moviegoers (adopters). The Bass forecasting model seems appropriate here.

These two films illustrate drastically different adoption patterns. Note that revenues for the independent studio film grow until the revenues peak in week 4 and then decline. For this film, much of the revenue is obviously due to word-of-mouth influence. In terms of the Bass model, the imitation factor dominates the innovation factor, and we expect $q > p$. However, for the summer blockbuster, revenues peak in week 1 and drop sharply afterward. The innovation factor dominates the imitation factor, and we expect $q < p$.

The forecasting model given in equation (16.18) can be incorporated into a nonlinear optimization problem to find the values of p, q, and m that give the best forecasts for a set of data. Assume that N periods of data are available. Let us denote the actual number of adopters (or a multiple of that number, such as sales) in period t as S_t for $t = 1, \ldots, N$. Then the forecast in each period and the corresponding forecast error E_t are defined by

$$F_t = \left(p + q[C_{t-1}/m]\right)(m - C_{t-1}) \text{ and } E_t = F_t - S_t$$

Notice that the forecast error is the difference between the forecast value F_t and the actual value S_t. It is common statistical practice to estimate the parameters p, q, and m by minimizing the sum of squared errors.

Doing so for the Bass forecasting model leads to the following nonlinear optimization problem:

$$\text{Min} \sum_{t=1}^{N} E_t^2 \qquad \textbf{(16.19)}$$

s.t.

$$F_t = (p + q\,[C_{t-1}/m])(m - C_{t-1}) \qquad t = 1, 2, \ldots, N \qquad \textbf{(16.20)}$$

$$E_t = F_t - S_t \qquad t = 1, 2, \ldots, N \qquad \textbf{(16.21)}$$

Note that the parameters of the Bass forecasting model are the decision variables in this nonlinear optimization model.

Because equations (16.19) and (16.20) both contain nonlinear terms, this model is a nonlinear minimization problem.

The data in Table 16.3 provide the revenue and cumulative revenues for the independent studio film in weeks 1–12. Using these data, the nonlinear model to estimate the parameters of the Bass forecasting model for the independent studio film is as follows:

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

$$
\begin{aligned}
\text{Min} \quad & E_1^2 + E_2^2 + \cdots + E_{12}^2 \\
\text{s.t.} \quad & F_1 = (p)m \\
& F_2 = [p + q(0.10/m)](m - 0.10) \\
& F_3 = [p + q(3.10/m)](m - 3.10) \\
& \vdots \\
& F_{12} = [p + q(34.85/m)](m - 34.85) \\
& E_1 = F_1 - 0.10 \\
& E_2 = F_2 - 3.00 \\
& \vdots \\
& E_{12} = F_{12} - 0.60
\end{aligned}
$$

The solutions to this nonlinear model and to a similar nonlinear model for the summer blockbuster are given in Table 16.4.

MODEL file
bass

The optimal forecasting parameter values given in Table 16.4 are intuitively appealing and consistent with Figure 16.13. For the independent studio film, which has the largest revenues in week 4, the value of the imitation parameter q is 0.49; this value is substantially larger than the innovation parameter $p = 0.074$. The film picks up momentum over time because of favorable word of mouth. After week 4, revenues decline as more and more of the potential market for the film has already seen it. Contrast these data with the summer blockbuster movie, which has a negative value of -0.018 for the imitation parameter q and an innovation parameter p of 0.49. The greatest number of adoptions is in week 1, and new adoptions decline afterward. Obviously the word-of-mouth influence is not favorable.

Table 16.3 Box Office Revenues and Cumulative Revenues in $ Millions for Independent Studio Film

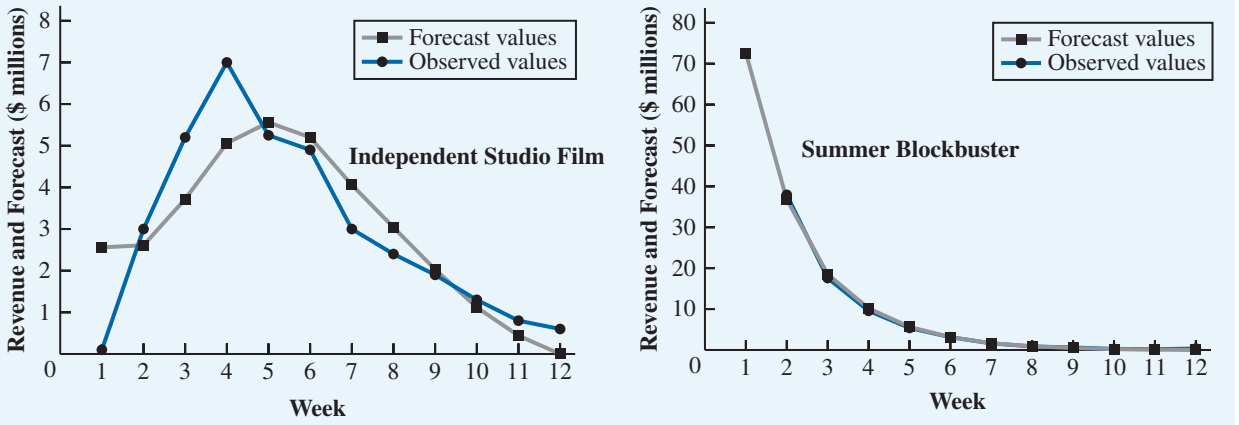
Week	Revenues S_t	Cumulative Revenues C_t
1	0.10	0.10
2	3.00	3.10
3	5.20	8.30
4	7.00	15.30
5	5.25	20.55
6	4.90	25.45
7	3.00	28.45
8	2.40	30.85
9	1.90	32.75
10	1.30	34.05
11	0.80	34.85
12	0.60	35.45

Table 16.4 Optimal Forecast Parameters for Independent Studio Film and Summer Blockbuster Movie

Parameter	Independent Studio Film	Summer Blockbuster
p	0.074	0.460
q	0.490	−0.018
m	34.850	149.540

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 16.14 Forecast and Actual Weekly Box Office Revenues for Independent Studio Film and Summer Blockbuster



In Figure 16.14, we show the forecast values based on the parameters in Table 16.4 and the observed values in the same graph. The Bass forecasting model does a good job of tracking revenue for the independent small-studio film. For the summer blockbuster, the Bass model does an outstanding job; it is virtually impossible to distinguish the forecast line from the actual adoption line.

You may wonder what good a forecasting model is if we must wait until after the adoption cycle is complete to estimate the parameters. One way to use the Bass forecasting model for a new product is to assume that sales of the new product will behave in a way that is similar to a previous product for which p and q values have been calculated and to subjectively estimate m, the potential market for the new product. For example, one might assume that box office receipts for movies next summer will behave similarly to box office receipts for movies last summer. Then the p and q values used for next summer's movies would be the p and q values calculated from the actual box office receipts last summer.

A second approach is to wait until several periods of data for the new product are available. For example, if five periods of data are available, the sales data for these five periods could be used to forecast demand for period 6. Then, after six periods of sales are observed, a forecast for period 7 is made. This method is often called a *rolling-horizon* approach.

Notes + Comments

The optimization model used to determine the parameter values for the Bass forecasting model is an example of a difficult nonlinear optimization problem. It is neither convex nor concave. For such models, local optima may give values that are much worse than the global optimum. We recommend using the Multistart option in Excel Solver when solving such problems.

16.6 Heuristic Optimization Using Excel's Evolutionary Method

Some optimization models of highly complex decisions can be difficult to model algebraically, but can be modeled in a spreadsheet using functions such as IF, COUNTIF, SUMIF, and XLOOKUP. Such cases are nonlinear and often cannot be solved to optimality by the *GRG Nonlinear* method in Excel Solver. In addition to the *Simplex LP* method for linear models and *GRG Nonlinear* method for nonlinear models, Excel Solver also has an alternative *Evolutionary* method. For complex nonlinear models, the *Evolutionary* method can often find a solution that is superior to the solution found using the *GRG Nonlinear* method.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In this section, we discuss a complex nonlinear model and the use of Excel's *Evolutionary* method to find a solution.

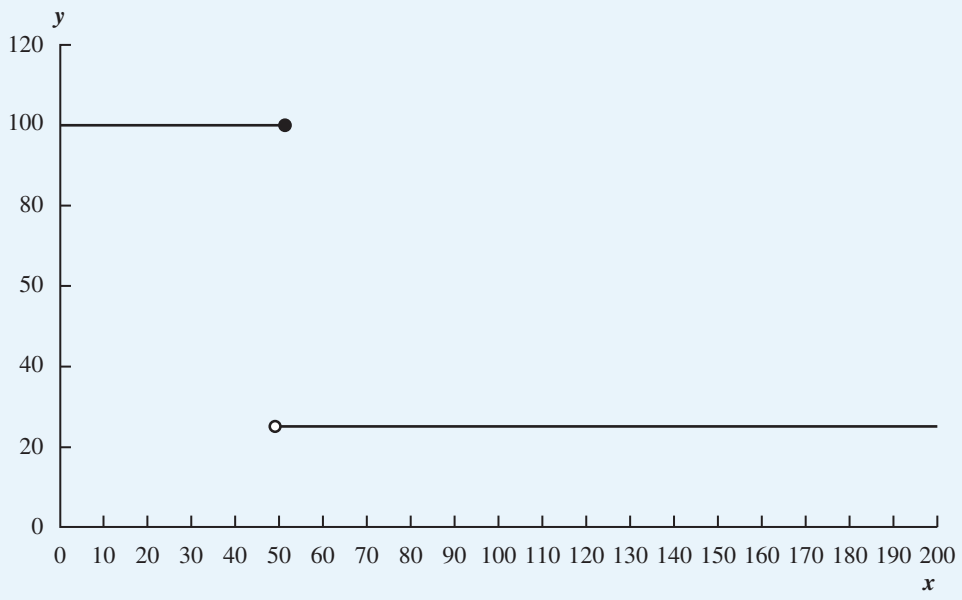
At its core, the *Evolutionary* method in Excel Solver is a sophisticated trial-and-error method for solving a complex model that involves non-smooth functions such as IF, COUNTIF, SUMIF, and XLOOKUP. For example, the function:

$$y = 100 \text{ if } x \leq 50$$
$$= 25 \text{ if } x > 50$$

Notice there is a discrete drop from a value of 100 at $x = 50$ to a value of 25 for x strictly larger than 50. Traditional nonlinear algorithms such as *GRG Nonlinear* can struggle with these non-smooth functions, because numerically it is difficult to project how the function is behaving when it is not smooth. Consequently, a quick and intelligent trial-and-error approach can often perform better on these types of problems. Trial-and-error approaches essentially sample trial points and use rules for generating new points to find good solutions. These types of trial-and-error methods are known as heuristic methods. A **heuristic method** is an approach that can quickly obtain a solution to an optimization problem, but this solution is not guaranteed to be optimal. The *Evolutionary* method in Excel Solver is a heuristic method.

The *Evolutionary* method derives its name from its "survival of the fittest" approach. At any point in the solution process, a population of solutions is stored. The stored population of solutions is then subject to "crossover" and "mutation." In the crossover process, a new trial solution is generated by determining the value of its decision variables from a combination or "crossover" of two solutions chosen from the population. In the mutation process, the values of some decision variables in a solution chosen from the population are randomly changed to generate a new trial solution. A new trial solution generated by crossover and/or mutation is then considered for inclusion in the population based on its "fitness." The fitness function can take into account the solution's feasibility and objective function value. Through crossover and mutation, the fitness of the population evolves (improves over time). The algorithm can be stopped based on run time (setting a maximum run time without any improvement in the best solution found thus far) or when the members of the population all have very similar fitness values.

Figure 16.15 A Nonsmooth Function



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Randomness in the mutation process helps ensure that a broad range of solutions is tried, but also means multiple runs of the *Evolutionary* method can end with different final solutions.

Figure 16.16 shows the Evolutionary Solver options dialog box and default parameter settings. A brief description of each follows.

Convergence—the Evolutionary method will stop if the population has a maximum difference in fitness that does not exceed the value of the convergence parameter. A larger value for the convergence parameter will result in a faster run, but will potentially result in a lower-quality solution.

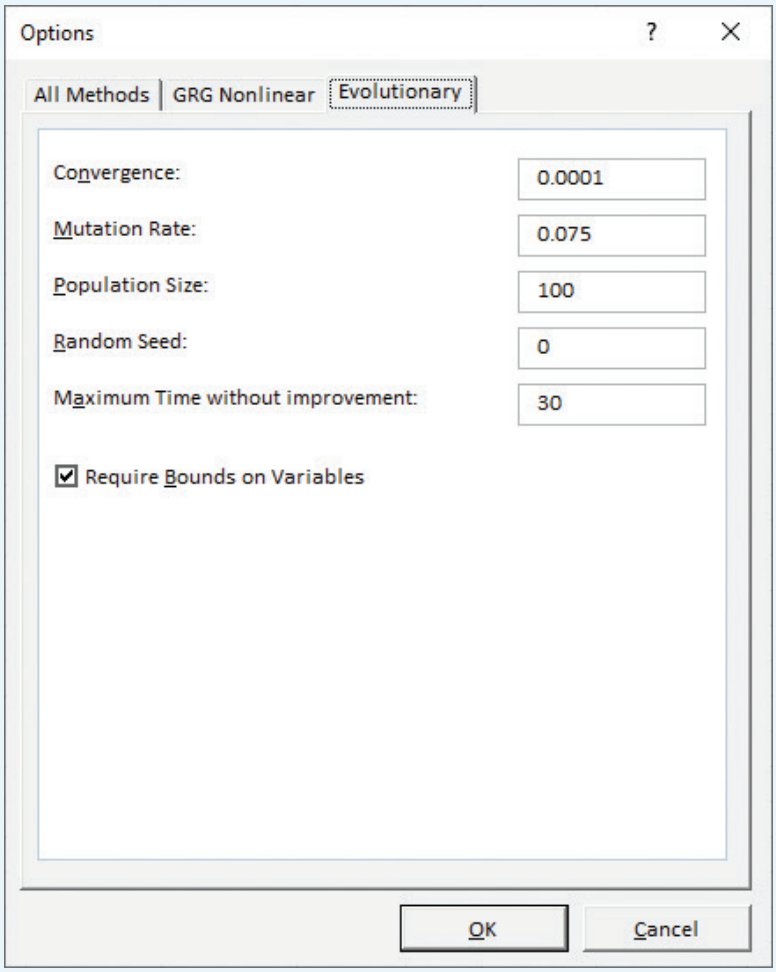
Mutation Rate—this is the probability that a solution in the population will be mutated. The higher the mutation rate, the more diverse the trial solutions.

Population Size—this is the number of solutions stored. The larger the population, the larger the set of candidate solutions considered for mutation.

Random Seed—this parameter is used to control the random numbers used to mutate the solutions. If a positive integer is entered, it will generate a set of random numbers that can be repeated if the same seed is used in a later run (so the analyst can replicate a solution when debugging a model or comparing model changes).

Maximum Time without Improvement—this is the maximum time the algorithm will run without improvement in the best solution.

Figure 16.16 The Excel Evolutionary Solver Options Dialog Box



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

We recommend starting with the default values of the parameters. If you find a solution that is not satisfactory when running the *Evolutionary* method, you might consider increasing the mutation rate and/or lowering the convergence rate. However, both of these changes will likely increase the run time. Next we demonstrate the use of the *Evolutionary* method on a pricing problem.

As an example of a complex model for which it is appropriate to use the *Evolutionary* method, let us consider the problem faced by Mount Airy College. To increase revenue from the sale of football tickets for its home games, Mount Airy College is considering using bundle pricing for the upcoming football season. **Bundle pricing** involves grouping items usually sold separately into bundles in various combinations to be sold at a price cheaper than if the items in the bundle were purchased separately. The idea is that the bundle price will entice some customers to buy items they would not normally buy, because they are bundled with other more-enticing items and offered at a cheaper price than if all items were purchased individually. The overall objective of bundle pricing is to maximize the revenue of the sale of the items; in the case of Mount Airy College, that is revenue from football ticket sales.

To help analyze their pricing problem, Mt. Airy College has conducted a survey of 75 potential ticket buyers and asked them what they would be willing to pay for individual home games this season against North Carolina State University, Army, and Duke University. For example, as shown in Figure 16.17, the first respondent is willing to pay $28 to attend the Mt. Airy versus NC State game, $28 to attend the Army game, and $49 to attend the Duke game. The other rows of data are interpreted in a similar manner for other respondents.

Figure 16.17 A Portion of the Pricing Data for Mt. Airy Football Tickets

	A	B	C	D
1	**Mt Airy Football**			
2				
3	**Data**			
4	Respondent	NC State	Army	Duke
5	1	$28.00	$28.00	$49.00
6	2	$70.00	$70.00	$0.00
7	3	$0.00	$7.00	$14.00
8	4	$0.00	$70.00	$0.00
9	5	$21.00	$35.00	$10.50
10	6	$7.00	$21.00	$0.00
11	7	$5.25	$7.00	$2.10
12	8	$21.00	$28.00	$28.00
13	9	$35.00	$49.00	$21.00
14	10	$21.00	$21.00	$21.00
15	11	$14.00	$35.00	$21.00
16	12	$14.00	$35.00	$14.00

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 16.5 The Seven Ticket Products and Proposed Prices for Mt. Airy Football Home Games

Individual Games		Two-Game Bundles		Three-Game Bundle	
NC State	$40	NC State + Army	$60	NC State + Army + Duke	$95
Army	$40	NC State + Duke	$60		
Duke	$40	Army + Duke	$60		

Individual game tickets and bundles of tickets for Mt. Airy are shown in Table 16.5, along with the proposed pricing strategy. Mt. Airy will offer single-game tickets, bundles of two games, and a bundle consisting of all three home games, a total of seven different products.

In bundle pricing, we need to ensure that it is never cheaper to buy a combination of individual products than to buy the corresponding bundled product. For example, to be an effective product, the three-game bundle price must always cost no more than purchasing the three games individually at their individual game prices. Likewise, it should not be cheaper to buy the bundle of NC State and Army and the individual Duke game than it is to buy the three-game bundle. Mt. Airy has also decided that no individual game ticket price can exceed $50, no two-game ticket price can exceed $80, and the three-game price cannot exceed $120. Finally, prices must be integer valued.

Let us define an algebraic model of this pricing problem. The model is to maximize ticket revenue while adhering to the pricing rules and the upper limits placed on the prices. If we define the following: NC State as game 1, Army as game 2, and Duke as game 3, we can use subscripts that define the combinations of games under consideration.

Let

X_1 = price of a ticket to the NC State game
X_2 = price of a ticket to the Army game
X_3 = price of a ticket to the Duke game
X_{12} = bundle price of a ticket to both the NC State and Army games
X_{13} = bundle price of a ticket to both the NC State and Duke games
X_{23} = bundle price of a ticket to both the Army and Duke games
X_{123} = bundle price of a ticket to all three games

The optimization model is as follows:

$$\text{Max Revenue}(X_1, X_2, X_3, X_{12}, X_{13}, X_{23}, X_{123}) \quad \textbf{(16.22)}$$

s.t.

$$X_{12} \leq X_1 + X_2 \quad \textbf{(16.23)}$$
$$X_{13} \leq X_1 + X_3 \quad \textbf{(16.24)}$$
$$X_{23} \leq X_2 + X_3 \quad \textbf{(16.25)}$$
$$X_{123} \leq X_1 + X_2 + X_3 \quad \textbf{(16.26)}$$
$$X_{123} \leq X_1 + X_{23} \quad \textbf{(16.27)}$$
$$X_{123} \leq X_2 + X_{13} \quad \textbf{(16.28)}$$
$$X_{123} \leq X_3 + X_{12} \quad \textbf{(16.29)}$$
$$X_1, X_2, X_3 \leq 50 \quad \textbf{(16.30)}$$
$$X_{12}, X_{13}, X_{23} \leq 80 \quad \textbf{(16.31)}$$
$$X_{123} \leq 120 \quad \textbf{(16.32)}$$
$$X_1, X_2, X_3, X_{12}, X_{13}, X_{23}, X_{123} \geq 0 \text{ and integer} \quad \textbf{(16.33)}$$

Let us first discuss the constraints of this model. Constraints (16.23)–(16.25) ensure that two-game bundle prices are no more than the sum of their corresponding two individual

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

game prices. Constraint (16.26) ensures that the three-game bundle price is no more than the than the sum of three individual game prices. Constraints (16.27)–(16.29) ensure that the three-game bundle price is no more than the price of a single game plus the two-game bundle. Constraints (16.30), (16.31), and (16.32) limit individual game prices, two-game bundle prices and the three-game bundle price to no more than $50, $80, and $120 respectively. Constraint (16.33) ensures that all prices are nonnegative and integer.

A complicating factor is how to calculate the objective function (16.22), which for now we have simply denoted as a function called *Revenue,* a function of the product prices. Pricing analysts have found that consumers purchase a product based on the product's **consumer surplus**, the difference between the value the consumer places on the product and its cost. Given the choice of several different alternatives for purchase, the consumer will purchase the product with the highest positive consumer surplus. Hence, for a given respondent, we can calculate revenue as follows:

- For a given set of prices for the seven products, calculate the consumer surplus.
- Find the maximum of the seven consumer surpluses.
- If the maximum consumer surplus is negative, the respondent will not purchase any of the products and revenue will be $0.
- If the maximum consumer surplus is nonnegative, the respondent will purchase the product that resulted in the maximum consumer surplus and the revenue will be the price of that product.

Based on this, we have constructed a model that will calculate revenue for ticket sales for Mt. Airy College. A partial view of the bundle pricing model is shown in Figure 16.18.

As partially shown in Figure 16.18, the cell range A87:D161 contains a copy of the pricing data from the survey of 75 potential ticket buyers that is partially shown in Figure 16.17. The shaded cells, E84 through K84, are the current prices proposed for products 1 through 7. The price of the individual games is $40 per game, the price of the three two-game bundles is set at $60, and the price of the three-game bundle is $95. In the cell range E87:K161, the consumer surplus is calculated for each respondent and product. For example, cell H87 contains the formula for the consumer surplus for respondent 1 for the two-game bundle consisting of games with NC State and Army:

$$=\text{B87} + \text{C87} - \text{H\$84} = 28.00 + 28.00 - 60.00 = -4.00$$

Hence, at the current prices for this respondent and bundle, there is a negative consumer surplus, making this an unattractive product for respondent 1.

The Excel MAX *function returns the maximum value found in a range of cells.*

In column L, the maximum consumer surplus is calculated for each respondent using the Excel MAX function. For example, for respondent 1, the MAX function is in cell L87 and is =MAX(E87:K87), and the maximum that is returned is 17 from cell I87. Note that there is a tie for the maximum (both cells I87 and J87 have values of 17). In the case of ties, the first match from the left is selected.

The Excel IF *function is discussed in Chapter 12.*

If the maximum consumer surplus in column L is nonnegative, we assume the respondent would buy the product that yields the maximum consumer surplus. The function in column M returns the product that gives the maximum surplus. If the maximum surplus is negative, there is no purchase and a zero is returned. The function used to calculate this for respondent 1 is found in cell M87 and is as follows:

$$=\text{IF(L87<0,0,MATCH(L87,E87:K87,0))}$$

If the value in L87 is less than 0, then the IF function returns the second argument (which is 0). If the value in L87 is greater than or equal to 0, then the IF function returns the value of the third argument, which is:

$$\text{MATCH(L87,E87:K87,0)}$$

The third argument of 0 in the MATCH function implies the search is for an *exact match* of the value in the first argument (the value in L87). The search is performed over the range

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

given in the second argument (E87:K87). The MATCH function returns the *position* of the matched value in the range of the second argument. In row 87 in Figure 16.18 the MATCH function returns the value of 5 in cell M87 because the value of 17.00 appears in the fifth position of the range in E87:K87.

As another example of this calculation, note that in row 89 the value in M89 is zero because the value in cell L89 is negative. Therefore, the function

$$=\text{IF(L89<0,0,MATCH(L89,E89:K89,0))}$$

returns a 0 because the value in L89 is −26.00, which is less than 0.

The function in column N calculates revenue from each respondent using an IF function with an XLOOKUP function. To get the revenue for a respondent, we need to get the price for the relevant product from the range of cells E84:K84 or return a 0 if no product will be purchased by the respondent. The following function is in cell N87 to calculate the revenue for respondent 1:

$$=\text{IF(M87=0,0,XLOOKUP(M87,\$E\$83:\$K\$83,\$E\$84:\$K\$84, N/A, −1,1))}$$

The IF function first tests if M87 = 0 (the first argument of the function). If the answer is yes, then a zero (the second argument) is returned. If the value in M87 is not zero, then the IF statement uses the XLOOKUP function:

MODEL file
mtairy

$$\text{XLOOKUP(M87,\$E\$83:\$K\$83,\$E\$84:\$K\$84, N/A, −1,1)}$$

The XLOOKUP function searches for the value in M87 in the range $E$83:$K$83 and returns the value in the same position in the range $E$84:$K$84. In this case, the value in

Figure 16.18 Partial View of Mt. Airy Football Bundle Pricing Model

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
81	**Model**																		
82	Total Revenue	$2,330.00																	
83				Product	1	2	3	4	5	6	7								
84				Price	$40.00	$40.00	$40.00	$60.00	$60.00	$60.00	$95.00								
85				Games	1	2	3	1,2	1, 3	2, 3	1, 2, 3								
86	Respondent	Game 1: NC State	Game 2: Army	Game 3: Duke	NC State	Army	Duke	NC State + Army	NC State + Duke	Army + Duke	NC State + Army + Duke	Max Consumer Suplus	Product Bought	Revenue		Games	Priced Separately	Price Bundled	Savings
87	1	$28.00	$28.00	$49.00	-12.00	-12.00	9.00	-4.00	17.00	17.00	10.00	17.00	5	$60.00		1 and 2	$80.00	$60.00	$20.00
88	2	$70.00	$70.00	$0.00	30.00	30.00	-40.00	80.00	10.00	10.00	45.00	80.00	4	$60.00		1 and 3	$80.00	$60.00	$20.00
89	3	$0.00	$7.00	$14.00	-40.00	-33.00	-26.00	-53.00	-46.00	-39.00	-74.00	-26.00	0	$0.00		2 and 3	$80.00	$60.00	$20.00
90	4	$0.00	$70.00	$0.00	-40.00	30.00	-40.00	10.00	-60.00	10.00	-25.00	30.00	2	$40.00		1,2, and 3	$120.00	$95.00	$25.00
91	5	$21.00	$35.00	$10.50	-19.00	-5.00	-29.50	-4.00	-28.50	-14.50	-28.50	-4.00	0	$0.00		3 and 4	$100.00	$95.00	$5.00
92	6	$7.00	$21.00	$0.00	-33.00	-19.00	-40.00	-32.00	-53.00	-39.00	-67.00	-19.00	0	$0.00		2 and 5	$100.00	$95.00	$5.00
93	7	$5.25	$7.00	$2.10	-34.75	-33.00	-37.90	-47.75	-52.65	-50.90	-80.65	-33.00	0	$0.00		1 and 6	$100.00	$95.00	$5.00
94	8	$21.00	$28.00	$28.00	-19.00	-12.00	-12.00	-11.00	-11.00	-4.00	-18.00	-4.00	0	$0.00					

	E	F	G	H	I	J	K	L
86	NC State	Army	Duke	NC State + Army	NC State + Duke	Duke + Army	NC State + Army + Duke	Max Consumer Surplus
87	=B87-E$84	=C87-F$84	=D87-G$84	=B87+C87-H$84	=B87+D87-I$84	=C87+D87-J$84	=SUM(B87:D87)-K$84	=MAX(E87:K87)
88	=B88-E$84	=C88-F$84	=D88-G$84	=B88+C88-H$84	=B88+D88-I$84	=C88+D88-J$84	=SUM(B88:D88)-K$84	=MAX(E88:K88)
89	=B89-E$84	=C89-F$84	=D89-G$84	=B89+C89-H$84	=B89+D89-I$84	=C89+D89-J$84	=SUM(B89:D89)-K$84	=MAX(E89:K89)
90	=B90-E$84	=C90-F$84	=D90-G$84	=B90+C90-H$84	=B90+D90-I$84	=C90+D90-J$84	=SUM(B90:D90)-K$84	=MAX(E90:K90)
91	=B91-E$84	=C91-F$84	=D91-G$84	=B91+C91-H$84	=B91+D91-I$84	=C91+D91-J$84	=SUM(B91:D91)-K$84	=MAX(E91:K91)
92	=B92-E$84	=C92-F$84	=D92-G$84	=B92+C92-H$84	=B92+D92-I$84	=C92+D92-J$84	=SUM(B92:D92)-K$84	=MAX(E92:K92)
93	=B93-E$84	=C93-F$84	=D93-G$84	=B93+C93-H$84	=B93+D93-I$84	=C93+D93-J$84	=SUM(B93:D93)-K$84	=MAX(E93:K93)
94	=B94-E$84	=C94-F$84	=D94-G$84	=B94+C94-H$84	=B94+D94-I$84	=C94+D94-J$84	=SUM(B94:D94)-K$84	=MAX(E94:K94)

	M	N	O	P	Q	R	S
86	Product Bought	Revenue		Products	Sum of Products Priced Separately	Price Bundled	Savings
87	=IF(L87<0,0,MATCH(L87,E87:K87,0))	=IF(M87=0,0,XLOOKUP(M87,$E$83:$K$83,$E$84:$K$84, N/A, -1,1))		1 and 2	=E84+F84	=H84	=Q87-R87
88	=IF(L88<0,0,MATCH(L88,E88:K88,0))	=IF(M88=0,0,XLOOKUP(M88,$E$83:$K$83,$E$84:$K$84, N/A, -1,1))		1 and 3	=E84+G84	=I84	=Q88-R88
89	=IF(L89<0,0,MATCH(L89,E89:K89,0))	=IF(M89=0,0,XLOOKUP(M89,$E$83:$K$83,$E$84:$K$84, N/A, -1,1))		2 and 3	=F84+G84	=J84	=Q89-R89
90	=IF(L90<0,0,MATCH(L90,E90:K90,0))	=IF(M90=0,0,XLOOKUP(M90,$E$83:$K$83,$E$84:$K$84, N/A, -1,1))		1,2, and 3	=E84+F84+G84	=K84	=Q90-R90
91	=IF(L91<0,0,MATCH(L91,E91:K91,0))	=IF(M91=0,0,XLOOKUP(M91,$E$83:$K$83,$E$84:$K$84, N/A, -1,1))		3 and 4	=G84+H84	=K84	=Q91-R91
92	=IF(L92<0,0,MATCH(L92,E92:K92,0))	=IF(M92=0,0,XLOOKUP(M92,$E$83:$K$83,$E$84:$K$84, N/A, -1,1))		2 and 5	=F84+I84	=K84	=Q92-R92
93	=IF(L93<0,0,MATCH(L93,E93:K93,0))	=IF(M93=0,0,XLOOKUP(M93,$E$83:$K$83,$E$84:$K$84, N/A, -1,1))		1 and 6	=E84+J84	=K84	=Q93-R93
94	=IF(L94<0,0,MATCH(L94,E94:K94,0))	=IF(M94=0,0,XLOOKUP(M94,$E$83:$K$83,$E$84:$K$84, N/A, -1,1))					

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The excel XLOOKUP *function is discussed in Chapter 12.*

M87 is 5, the exact match of 5 is found in I83 and the function returns the value in the fifth position in the row below which is $60 found in cell I84. Note that the last three arguments of the XLOOKUP function (N/A, −1, 1) ensure the following: the N/A results in "N/A" being returned if no match is found; -1 means the search is for an exact match and the 1 means start at the beginning of the specified search range.

In columns P through S, we give the bundled price and the price of the items if purchased separately, along with the savings for choosing the bundled product.

In cell B82, total revenue from all 75 respondents is calculated using the function =SUM(N87:161). We see that for the proposed pricing strategy, total ticket revenue is $2,330.00.

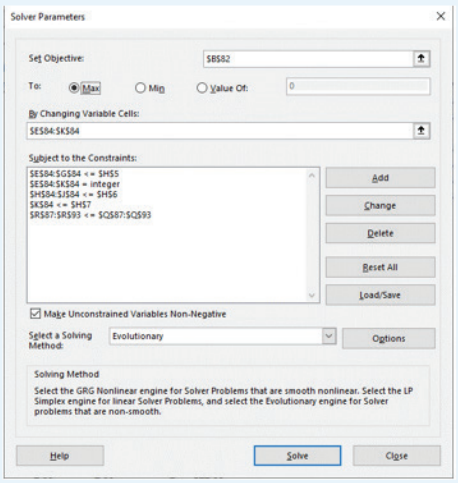
Can Mt. Airy do better than the proposed pricing strategy shown in Figure 16.18? The objective of Mt. Airy is to find the price of each product that maximizes total revenue over all the respondents. Then, the optimal prices will be used for the upcoming season under the assumption that the sample consumer data were representative of the population of potential customers.

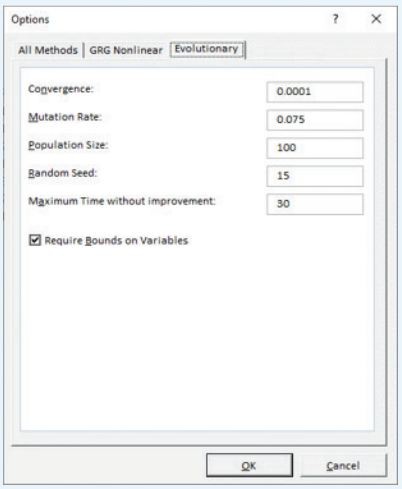
We ran the *Evolutionary* Solver using the default parameters (see Figure 16.16) but with a random seed of 15 and the proposed pricing as shown in Table 16.5 as starting values for the prices. The Solver model is also shown in Figure 16.19. The run of the *Evolutionary* method stopped with the message "Solver Cannot Improve the Current Solution." The best solution found is shown in Figure 16.19.

The best solution found by the *Evolutionary* method generates revenue of $3,339. Since the Evolutionary algorithm is a heuristic, we are not guaranteed that this is the optimal solution. However, we do know that this solution provides 43% more revenue than the proposed pricing strategy ($3,339 versus $2,330) for this sample data. The pricing strategy found by the Evolutionary methods is shown in Table 16.6.

Figure 16.19 The Mt. Airy Football Pricing Model and Best Solution Found

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
81	**Model**																		
82	Total Revenue	$3,339.00																	
83				Product	1	2	3	4	5	6	7								
84				Price	$45.00	$35.00	$44.00	$70.00	$70.00	$70.00	$77.00								
85				Games	1	2	3	1,2	1, 3	2, 3	1, 2, 3								
86	Respondent	Game 1: NC State	Game 2: Army	Game 3: Duke	NC State	Army	Duke	NC State + Army	NC State + Duke	Army + Duke	NC State + Army + Duke	Max Consumer Suplus	Product Bought	Revenue		Games	Priced Separately	Price Bundled	Savings
87	1	$28.00	$28.00	$49.00	-17.00	-7.00	5.00	-14.00	7.00	7.00	28.00	28.00	7	$77.00		1 and 2	$80.00	$70.00	$10.00
88	2	$70.00	$70.00	$0.00	25.00	35.00	-44.00	70.00	0.00	0.00	63.00	70.00	4	$70.00		1 and 3	$89.00	$70.00	$19.00
89	3	$0.00	$7.00	$14.00	-45.00	-28.00	-30.00	-63.00	-56.00	-49.00	-56.00	-28.00	0	$0.00		2 and 3	$79.00	$70.00	$9.00
90	4	$0.00	$70.00	$0.00	-45.00	35.00	-44.00	0.00	-70.00	0.00	-7.00	35.00	2	$35.00		1,2, and 3	$124.00	$77.00	$47.00
91	5	$21.00	$35.00	$10.50	-24.00	0.00	-33.50	-14.00	-38.50	-24.50	-10.50	0.00	2	$35.00		3 and 4	$114.00	$77.00	$37.00
92	6	$7.00	$21.00	$0.00	-38.00	-14.00	-44.00	-42.00	-63.00	-49.00	-49.00	-14.00	0	$0.00		2 and 5	$105.00	$77.00	$28.00
93	7	$5.25	$7.00	$2.10	-39.75	-28.00	-41.90	-57.75	-62.65	-60.90	-62.65	-28.00	0	$0.00		1 and 6	$115.00	$77.00	$38.00
94	8	$21.00	$28.00	$28.00	-24.00	-7.00	-16.00	-21.00	-21.00	-14.00	0.00	0.00	7	$77.00					





Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 16.6 The Product Prices Found by the Evolutionary Method for Mt. Airy Football Home Games

Individual Games		Two-Game Bundles		Three-Game Bundle	
NC State	$45	NC State + Army	$70	NC State + Army + Duke	$77
Army	$35	NC State + Duke	$70		
Duke	$44	Army + Duke	$70		

Notes + Comments

1. Because the Evolutionary algorithm is probabilistic, solving the same model multiple times (even from the same starting point) can result in different final solutions. However, by setting the Evolutionary option **Random Seed** to an integer value other than the default value of 0, you can fix the set of random numbers used in the solution process. Then, starting from the same solution, the same final solution should be achieved in multiple runs. This can be useful in debugging a model.
2. The Evolutionary method will pause when it hits its limit on solution time, time without improvement, or limit on the number of feasible solutions tried. A dialog box will allow you to stop and report the best solution found thus far or continue the search process.
3. At any time in a run of the Evolutionary method, you may press the escape key (Esc) on your keyboard to pause the process. As is the case when the method hits one of its limits, a dialog box will allow you to stop and report the best solution found thus far or continue the search process.

Summary

In this chapter, we introduced nonlinear optimization models. A nonlinear optimization model is a model with at least one nonlinear term in either a constraint or the objective function. Because so many applications of business analytics involve nonlinear functions, allowing nonlinear terms greatly increases the number of important applications that can be modeled as an optimization problem. Numerous problems in portfolio optimization, option pricing, marketing, economics, facility location, forecasting, and scheduling lend themselves to nonlinear models.

Unfortunately, nonlinear optimization models are not as easy to solve as linear optimization models, or even integer linear optimization models. As a rule of thumb, if a problem can be modeled realistically as a linear or integer linear problem, then it is probably best to do so. Many nonlinear formulations have local optima that are not globally optimal. Because most nonlinear optimization software terminates with a local optimum, the solution returned by the software may not be the best solution available. However, as discussed in this chapter, numerous important classes of optimization problems, such as the Markowitz portfolio models, are convex optimization problems. For a convex optimization problem, a local optimum is also the global optimum. Additionally, the development of software for solving (nonconvex) nonlinear optimization problems that find globally optimal solutions is proceeding at a rapid rate. When using Excel Solver for nonlinear optimization, we recommend using the Multistart option. For very complex and nonlinear models, heuristic approaches such as Excel's Evolutionary method are useful for finding good solutions.

Glossary

Bundle pricing Grouping items usually sold separately into bundles in various combinations to be sold at a price cheaper than if the items in the bundle were purchased separately.
Concave function A function that is bowl-shaped down: For example, the functions $f(x) = -5x^2 - 5x$ and $f(x, y) = -x^2 - 11y^2$ are concave functions.
Consumer surplus The difference between the value the consumer places on the product and its cost.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Convex function A function that is bowl-shaped up: For example, the functions $f(x) = x^2 - 5x$ and $f(x, y) = x^2 + 5y^2$ are convex functions.
Efficient frontier A set of points defining the minimum possible risk (measured by portfolio variance) for a set of return values.
Global maximum A feasible solution is a global maximum if there are no other feasible points with a larger objective function value in the entire feasible region. A global maximum is also a local maximum.
Global minimum A feasible solution is a global minimum if there are no other feasible points with a smaller objective function value in the entire feasible region. A global minimum is also a local minimum.
Global optimum A feasible solution is a global optimum if there are no other feasible points with a better objective function value in the entire feasible region. A global optimum may be either a global maximum or a global minimum.
Heuristic method An approach that can quickly obtain a solution to an optimization problem that is not guaranteed to be optimal.
Lagrangian multiplier The shadow price for a constraint in a nonlinear problem, that is, the rate of change of the objective function with respect to the right-hand side of a constraint.
Local maximum A feasible solution is a local maximum if there are no other feasible solutions with a larger objective function value in the immediate neighborhood.
Local minimum A feasible solution is a local minimum if there are no other feasible solutions with a smaller objective function value in the immediate neighborhood.
Local optimum A feasible solution is a local optimum if there are no other feasible solutions with a better objective function value in the immediate neighborhood. A local optimum may be either a local maximum or a local minimum.
Markowitz mean-variance portfolio model An optimization model used to construct a portfolio that minimizes risk subject to a constraint requiring a minimum level of return.
Nonlinear optimization problem An optimization problem that contains at least one nonlinear term in the objective function or a constraint.
Quadratic function A nonlinear function with terms to the power of two.
Reduced gradient The value associated with a variable in a nonlinear model that is analogous to the reduced cost in a linear model; the shadow price of a binding simple lower or upper bound on the decision variable.

Problems

1. **Media Planning.** GreenLawns provides a lawn fertilizing and weed control service. The company is adding a special aeration treatment as a low-cost extra service option that it hopes will help attract new customers. Management is planning to promote this new service in two media: radio and direct-mail advertising. A media budget of $3,000 is available for this promotional campaign. Based on past experience in promoting its other services, GreenLawns has obtained the following estimate of the relationship between sales and the amount spent on promotion in these two media:

$$S = -2R^2 - 10M^2 - 8RM + 18R + 34M$$

where

S = total sales in thousands of dollars
R = thousands of dollars spent on radio advertising
M = thousands of dollars spent on direct-mail advertising

GreenLawns would like to develop a promotional strategy that will lead to maximum sales subject to the restriction provided by the media budget. **LO 1, 2**

a. What is the value of sales if $2,000 is spent on radio advertising and $1,000 is spent on direct-mail advertising?
b. Formulate an optimization problem that can be solved to maximize sales subject to the media budget of spending no more than $3,000 on total advertising.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

c. Determine the optimal amount to spend on radio and direct-mail advertising. How much in sales will be generated?

2. **Estimating Economic Output.** The Cobb-Douglas production function is a classic model from economics used to model output as a function of capital and labor. It has the form

$$f(L,C) = c_0 L^{c_1} C^{c_2}$$

where c_0, c_1, and c_2 are constants. The variable L represents the units of input of labor, and the variable C represents the units of input of capital. **LO 1, 2, 3**

a. In this example, assume $c_0 = 5$, $c_1 = 0.25$, and $c_2 = 0.75$. Assume each unit of labor costs \$25 and each unit of capital costs \$75. With \$75,000 available in the budget, develop an optimization model to determine how the budgeted amount should be allocated between capital and labor in order to maximize output.

b. Find the optimal solution to the model you formulated in part (a). *Hint:* When using Excel Solver, use the Multistart option with bounds $0 \le L \le 3{,}000$ and $0 \le C \le 1{,}000$.

3. **Steel Production Planning.** Let S represent the amount of steel produced (in tons). Steel production is related to the amount of labor used (L) and the amount of capital used (C) by the following function:

$$S = 20\, L^{0.30} C^{0.70}$$

In this formula L represents the units of labor input and C the units of capital input. Each unit of labor costs \$50, and each unit of capital costs \$100. **LO 1, 2, 3**

a. Formulate an optimization problem that will determine how much labor and capital are needed to produce 50,000 tons of steel at minimum cost.

b. Solve the optimization problem you formulated in part (a). *Hint:* When using Excel Solver, start with an initial $L > 0$ and $C > 0$.

4. **Production Planning for Two Products.** The profit function for two products is

$$\text{Profit} = -3x_1^2 + 42x_1 - 3x_2^2 + 48x_2 + 700$$

where x_1 represents units of production of product 1 and x_2 represents units of production of product 2. Producing one unit of product 1 requires 4 labor-hours and producing one unit of product 2 requires 6 labor-hours. Currently, 24 labor-hours are available. The cost of labor-hours is already factored into the profit function, but it is possible to schedule overtime at a premium of \$5 per hour. **LO 1, 2**

a. Formulate an optimization problem that can be used to find the optimal production quantity of products 1 and 2 and the optimal number of overtime hours to schedule.

b. Solve the optimization model you formulated in part (a). How much should be produced and how many overtime hours should be scheduled?

5. **Pricing Cameras.** Jim's Camera shop sells two high-end cameras, the Sky Eagle and Horizon. The demands for these two cameras are as follows: D_S = demand for the Sky Eagle, P_S is the selling price of the Sky Eagle, D_H is the demand for the Horizon, and P_H is the selling price of the Horizon.

$$\begin{aligned} D_S &= 222 - 0.60P_S + 0.35P_H \\ D_H &= 270 + 0.10P_S - 0.64P_H \end{aligned}$$

The store wishes to determine the selling price that maximizes revenue for these two products. Develop the revenue function for these two models, and find the prices that maximize revenue. **LO 1, 2**

6. **Baseball Glove Production Planning.** Heller Manufacturing has two production facilities that manufacture baseball gloves. Production costs at the two facilities differ because of varying labor rates, local property taxes, type of equipment, capacity, and so on. The Dayton plant has weekly costs that can be expressed as a function of the number of gloves produced:

$$TCD(X) = X^2 - X + 5$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

where X is the weekly production volume in thousands of units, and $TCD(X)$ is the cost in thousands of dollars. The Hamilton plant's weekly production costs are given by

$$TCH(Y) = Y^2 - 2Y + 3$$

where Y is the weekly production volume in thousands of units, and $TCH(Y)$ is the cost in thousands of dollars. Heller Manufacturing would like to produce 8,000 gloves per week at the lowest possible cost. **LO 1, 2**

a. Formulate a mathematical model that can be used to determine the optimal number of gloves to produce each week at each facility.
b. Solve the optimization model to determine the optimal number of gloves to produce at each facility.

7. **Exponential Smoothing.** Many forecasting models use parameters that are estimated using nonlinear optimization. A good example is the Bass model introduced in this chapter. Another example is the exponential smoothing forecasting model which is a common forecasting model used in practice. For instance, the basic exponential smoothing model for forecasting sales is

The exponential smoothing model is described in more detail in Chapter 9.

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t$$

where

$\hat{y}_{t+1}$ = forecast of sales for period $t + 1$
y_t = actual sales for period t
$\hat{y}_t$ = forecast of sales for period t
α = smoothing constant, $0 \leq \alpha \leq 1$

This model is used recursively; the forecast for time period $t + 1$ is based on the forecast for period t, $\hat{y}_t$, the observed value of sales in period t, y_t, and the smoothing parameter α. The use of this model to forecast sales for 12 months is illustrated in the following table with the smoothing constant $\alpha = 0.3$. The forecast errors, $y_t - \hat{y}_t$, are calculated in the fourth column. The value of a is often chosen by minimizing the sum of squared forecast errors. The last column of the table shows the square of the forecast error and the sum of squared forecast errors.

In using exponential smoothing models, one tries to choose the value of α that provides the best forecasts. **LO 1, 2**

a. The file *expsmoooth* contains the observed data shown here. Construct this table using the formula above. Note that we set the forecast in period 1 to the observed in period 1 to get started $(\hat{y}_1 = y_1 = 17)$, then the formula above for $\hat{y}_{t+1}$ is used starting in period 2. Make sure to have a single cell corresponding to α in your spreadsheet model. After confirming the values in the table below with $\alpha = 0.3$, try different values of α to see if you can get a smaller sum of squared forecast errors.

DATA*file*
expsmooth

Week (t)	Observed Value (y_t)	Forecast ($\hat{y}_t$)	Forecast Error ($y_t - \hat{y}_t$)	Squared Forecast Error $(y_t - \hat{y}_t)^2$
1	17	17.00	0.00	0.00
2	21	17.00	4.00	16.00
3	19	18.20	0.80	0.64
4	23	18.44	4.56	20.79
5	18	19.81	−1.81	3.27
6	16	19.27	−3.27	10.66
7	20	18.29	1.71	2.94
8	18	18.80	−0.80	0.64
9	22	18.56	3.44	11.83
10	20	19.59	0.41	0.17
11	15	19.71	−4.71	22.23
12	22	18.30	3.70	13.69
				SUM = 102.86

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

b. Use Excel Solver to find the value of α that minimizes the sum of squared forecast errors.

8. **Chair Manufacturing.** Andalus Furniture Company has two manufacturing plants, one at Aynor and another at Spartanburg. The cost in dollars of producing a kitchen chair at each of the two plants is given here. The cost of producing Q_1 chairs at Aynor is

$$75Q_1 + 5Q_1^2 + 100$$

and the cost of producing Q_2 kitchen chairs at Spartanburg is

$$25Q_2 + 2.5Q_2^2 + 150$$

Andalus needs to manufacture a total of 40 kitchen chairs to meet an order just received. How many chairs should be made at Aynor, and how many should be made at Spartanburg in order to minimize total production cost? **LO 1, 2**

9. **Economic Order Quantity.** The economic order quantity (EOQ) model is a classical model used for controlling inventory and satisfying demand. Costs included in the model are holding cost per unit, ordering cost, and the cost of goods ordered. The assumptions for that model are that only a single item is considered, that the entire quantity ordered arrives at one time, that the demand for the item is constant over time, and that no shortages are allowed.

Suppose we relax the first assumption and allow for multiple items that are independent except for a restriction on the amount of space available to store the products. The following model describes this situation:

Let

D_j = annual demand for item j
C_j = unit cost of item j
S_j = cost per order placed for item j
w_j = space required for item j
W = the maximum amount of space available for all goods
i = inventory carrying charge as a percentage of the cost per unit

The decision variables are Q_j, the amount of item j to order. The model is

$$\text{Minimize } \sum_{j=1}^{N}\left[C_jD_j + \frac{S_jD_j}{Q_j} + iC_j\frac{Q_j}{2}\right]$$

$$\text{s.t} \quad \sum_{j=1}^{N} w_jQ_j \le W$$

$$Q_j \ge 0 \qquad j = 1, 2, \ldots, N$$

In the objective function, the first term is the annual cost of goods, the second is the annual ordering cost (D_j/Q_j is the number of orders), and the last term is the annual inventory holding cost ($Q_i/2$ is the average amount of inventory).

Construct and solve a nonlinear optimization model for the following data. **LO 1, 2**

	Item 1	Item 2	Item 3
Annual Demand	2,000	2,000	1,000
Item Cost ($)	100	50	80
Order Cost ($)	150	135	125
Space Required (sq. feet)	50	25	40

$W = 5000$
$i = 0.20$

10. **Product Pricing.** Phillips Inc. produces two distinct products, A and B. The products do not compete with each other in the marketplace; that is, neither cost, price, nor demand for one product will impact the demand for the other. Phillips' analysts have

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

collected data on the effects of advertising on profits. These data suggest that, although higher advertising correlates with higher profits, the marginal increase in profits diminishes at higher advertising levels, particularly for product B. Analysts have estimated the following functions:

$$\text{Annual profit for product A} = 1.2712LN(X_A) + 17.414$$

$$\text{Annual profit for product B} = 0.3970LN(X_B) + 16.109$$

where X_A and X_B are the advertising amount allocated to products A and B, respectively, in thousands of dollars, profit is in millions of dollars, and *LN* is the natural logarithm function. The advertising budget is $500,000, and management has dictated that at least $50,000 must be allocated to each of the two products. *Hint:* To compute a natural logarithm for the value *X* in Excel, use the formula =*LN*(*X*). For Solver to find an answer, you also need to start with decision variable values greater than 0 in this problem. **LO 1, 2**

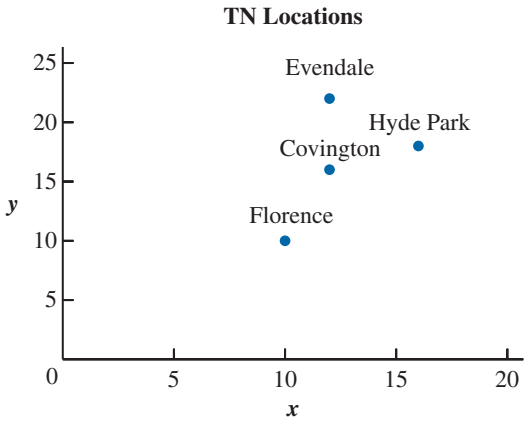
a. Build an optimization model that will prescribe how Phillips should allocate its marketing budget to maximize profit.
b. Solve the model you constructed in part (a) using Excel Solver.

11. **Tool Bin Location.** Let us consider again the data from the LaRosa tool bin location problem discussed in Section 16.3. **LO 1, 2**

DATA *file* larosademand

a. Suppose we know the average number of daily trips made to the tool bin from each production station. The average number of trips per day are 12 for fabrication, 24 for Paint, 13 for Subassembly 1, 7 for Subassembly 2, and 17 for Assembly. It seems as though we would want the tool bin closer to those stations with high average numbers of trips. Develop a new unconstrained model that minimizes the sum of the *demand-weighted distance* defined as the product of the demand (measured in number of trips) and the distance to the station.
b. Solve the model you developed in part (a). Comment on the differences between the unweighted distance solution given in Section 16.3 and the demand-weighted solution.

12. **Cell Tower Location.** TN Communications provides cellular telephone services. The company is planning to expand into the Cincinnati area and is trying to determine the best location for its transmission tower. The tower transmits over a radius of 10 miles. The locations that must be reached by this tower are shown in the following figure.



	x	y
Florence	10	10
Covington	12	16
Hyde Park	16	18
Evendale	12	22

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

TN Communications would like to find the tower location that reaches each of these cities and minimizes the sum of the distances to all locations from the new tower. **LO 1, 2**

a. Formulate and solve a model to find the optimal location.
b. Formulate and solve a model that minimizes the maximum distance from the transmission tower location to the city locations.

13. **Wedding Planning.** The "straight-line" distance between two cities in the United States can be approximated by the following formula, where lat_1 and $long_1$ are the latitude and longitude of city 1 and lat_2 and $long_2$ are the latitude and longitude of city 2, and the value 69 corresponds to the number of miles in a degree of latitude or longitude:

$$69\sqrt{(lat_1 - lat_2)^2 + (long_1 - long_2)^2}$$

DATA *file* wedding

Ted's daughter is getting married, and he is inviting relatives from 15 different locations in the United States. The file *wedding* gives the longitude, latitude, and number of relatives in each of the 15 locations. Ted would like to find a wedding location that minimizes the demand-weighted distance, where demand is the number of relatives at each location. Assuming that the wedding can occur anywhere, find the latitude and longitude of the optimal location. *Hint:* Notice that all longitude values given for this problem are negative. Make sure that you do *not* check the option for **Make Unconstrained Variables Non-Negative** in Solver. **LO 1, 2**

14. **Markowitz Portfolio Model**. Consider the stock return scenarios for Apple Computer (APPL), Advanced Micro Devices (AMD), and Oracle Corporation (ORCL) shown in the following table: **LO 1, 2, 4, 5, 6**

DATA *file* stockreturn1

	1	2	3	4	5	6	7	8
APPL	−39.80	10.10	124.90	151.80	−58.30	14.30	−41.90	57.10
AMD	−42.50	13.60	56.90	36.70	−34.80	−67.40	183.60	6.30
ORCL	−10.20	137.90	170.60	16.60	−40.70	−30.30	15.20	−0.60

a. Develop the Markowitz portfolio model for these data with a required expected return of 25%. Assume that the eight scenarios are equally likely to occur.
b. Solve the model developed in part (a).
c. Vary the required return in 1% increments from 25% to 30% and plot the efficient frontier.

15. **Alternative Markowitz Portfolio Model.** A second version of the Markowitz portfolio model maximizes expected return subject to a constraint that the variance of the portfolio must be less than or equal to some specified amount. Consider again the Hauck Financial Service data given in Section 16.4. **LO 1, 2, 4**

DATA *file* hauckdata

	Annual Return (%)				
Mutual Fund	Year 1	Year 2	Year 3	Year 4	Year 5
Foreign Stock	10.06	13.12	13.47	45.42	−21.93
Intermediate-Term Bond	17.64	3.25	7.51	−1.33	7.36
Large-Cap Growth	32.41	18.71	33.28	41.46	−23.26
Large-Cap Value	32.36	20.61	12.93	7.06	−5.37
Small-Cap Growth	33.44	19.40	3.85	58.68	−9.02
Small-Cap Value	24.56	25.32	−6.70	5.43	17.31

a. Construct this version of the Markowitz model for a maximum variance of 30.
b. Solve the model developed in part (a).

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

16. **Maximizing Minimum Return of Investment.** Reconsider the data in Problem 15. Construct a model that maximizes the minimum return achieved over the five scenarios provided. Solve your model to find the optimal portfolio. **LO 1, 2, 4**

17. **Portfolio Optimization.** Consider the following stock return data, **LO 1, 2, 4, 5, 6**

DATA*file* stockreturn2

	1	2	3	4	5	6
Stock 1	0.300	0.103	0.216	−0.046	−0.071	0.056
Stock 2	0.225	0.290	0.216	−0.272	0.144	0.107
Stock 3	0.149	0.260	0.419	−0.078	0.169	−0.035
	7	**8**	**9**	**10**	**11**	**12**
Stock 1	0.038	0.089	0.090	0.083	0.035	0.176
Stock 2	0.321	0.305	0.195	0.390	−0.072	0.715
Stock 3	0.133	0.732	0.021	0.131	0.006	0.908

a. Construct the Markowitz portfolio model using a required expected return of 15%. Assume that the 12 scenarios are equally likely to occur.
b. Solve the model using Excel Solver.
c. Solve the model for various values of required expected return and plot the efficient frontier.

18. **Matching the S&P 500 Return.** Let us consider again the investment data from Hauck Financial Services used in Section 16.4 to illustrate the Markowitz portfolio model. The data are shown below, along with the return of the S&P 500 Index. Hauck would like to create a portfolio using the funds listed, so that the resulting portfolio matches the return of the S&P 500 index as closely as possible. **LO 1, 2**

DATA*file* hauck500

Mutual Fund	Year 1	Year 2	Year 3	Year 4	Year 5
Foreign Stock	10.060	13.120	13.470	45.420	−21.930
Intermediate-Term Bond	17.640	3.250	7.510	−1.330	7.360
Large-Cap Growth	32.410	18.710	33.280	41.460	−23.260
Large-Cap Value	32.360	20.610	12.930	7.060	−5.370
Small-Cap Growth	33.440	19.400	3.850	58.680	−9.020
Small-Cap Value	24.560	25.320	−6.700	5.430	17.310
S&P 500 Return	25.000	20.000	8.000	30.000	−10.000

a. Develop an optimization model that will give the fraction of the portfolio to invest in each of the funds so that the return of the resulting portfolio matches as closely as possible the return of the S&P 500 Index. *Hint:* Minimize the sum of the squared deviations between the portfolio's return and the S&P 500 Index return for each year in the data set.
b. Solve the model developed in part (a).

19. **Semivariance as a Measure of Risk.** As discussed in Section 16.4, the Markowitz model uses the variance of the portfolio as the measure of risk. However, variance includes deviations both below and above the mean return. Semivariance includes only deviations below the mean and is considered by many to be a better measure of risk. **LO 1, 2, 4**

DATA*file* hauckdata

a. Develop a model that minimizes semivariance for the Hauck Financial data given in the file *hauckdata* with a required return of 10%. *Hint:* Modify model

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(16.8)–(16.17). Define a variable d_s for each scenario and let $d_s \geq \overline{R} - R_s$ with $d_s \geq 0$. Then make the objective function: Min $\frac{1}{5}\sum_{s=1}^{5} d_s^2$.

b. Solve the model you developed in part (a) with a required expected return of 10%.

20. **Constructing an Efficient Frontier.** Refer to Problem 15. Use the model developed there to construct an efficient frontier by varying the maximum allowable variance from 20 to 60 in increments of 5 and solving for the maximum return for each. Plot the efficient frontier and compare it to Figure 16.12. **LO 5, 6**

21. **Box Office Revenues.** The weekly box office revenues (in $ millions) for the summer blockbuster movie discussed in Section 16.5 follow. Use these data in the Bass forecasting model given by equations (16.19)–(16.21) to estimate the parameters p, q, and m.

DATA*file* blockbuster

Week	Revenues
1	72.39
2	37.93
3	17.58
4	9.57
5	5.39
6	3.13
7	1.62
8	0.87
9	0.61
10	0.26
11	0.19
12	0.35

The Bass forecasting model is a good example of a difficult-to-solve nonlinear program, and the answer you get may be a local optimum that is not nearly as good as the result given in Table 16.4. Solve the model using Excel Solver with the Multistart option, and see whether you can duplicate the results in Table 16.4. Use a lower bound of -1 and an upper bound of 1 on both p and q. Use a lower bound of 100 and an upper bound of 1,000 on m. **LO 1, 2, 3**

22. **Apparel Pricing.** A women's clothing retail chain has collected data on pricing and sales over the last five years at its flagship store in Charleston, SC. These data were used to estimate a regression equation that relates price to demand. The following estimated equation relates demand to the price for summer dresses:

$$Y = 1{,}000 - 1.89p$$

where Y = the demand for summer dresses and p = the price per dress.
Summer dresses cost $210. The data also show that when a summer dress is sold, on average one pair of shoes and one purse are sold with the dress. The profit on a pair of shoes is $18 and the profit on a purse is $26. **LO 1, 2**

a. What is the profit-maximizing price for dresses, ignoring the profit associated with the accompanying shoe and purse?
b. What is the profit-maximizing price for dresses taking into account the accompanying shoe and purse purchases?
c. Discuss the difference in prices obtained in parts (a) and (b).

23. **Calculating Stock Returns.** The purpose of this exercise is to learn how to calculate stock returns for portfolio models using actual stock price data. Ten years worth of

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

stock price data are included in the file *stockreturns* for Apple Computer (AAPL), Advanced Micro Devices (AMD), and Oracle Corporation (ORCL). These data contain closing prices that arc adjusted for stock dividends and splits.

Using these stock price data, we now want to calculate the annual returns for each stock for the Years 1 through 9. Returns are often calculated using continuous compounding. If the stock prices are adjusted for splits and stock dividends, then the price of stock i in period $t + 1$, $p_{i,t+1}$, is given by

$$P_{i,t+1} = P_{i,t}e^{r_{i,t}}$$

where $p_{i,t}$ is the price of stock i in period t and $r_{i,t}$ is the return on stock i in period t. This calculation assumes no cash dividends were paid, which is true of Apple Computer, Advanced Micro Devices, and Oracle Corporation for the years included here. Solving the equation $p_{i,t+1} = p_{i,t}e^{r_{i,t}}$ for the return on stock i in period t gives

$$r_{i,t} = \ln\left(\frac{p_{i,t+1}}{p_{i,t}}\right)$$

For example, the Apple Computer adjusted closing price in Year 9 was $38.45. The closing price in Year 10 was $75.51. Thus, the continuously compounded return for Apple Computer from Year 9 to Year 10 is

$$\ln(75.51/38.45) = 0.6749$$

We use this calculation as our estimate of the annual return for Apple Computer for Year 9.

Stock price data can be downloaded from many sites on the Internet including Yahoo! Finance, Google Finance, and others.

Take the closing stock prices shown in the following table and calculate the annual returns for Years 1 through 9 for AAPL, AMD, and ORCL using $r_{i,t} = \ln(p_{i,t+1}/p_{i,t})$. If you calculate the returns properly, your results should appear as in the final three columns of the following table. **LO 4**

DATA *file*
stockreturns

Year	AAPL Adj· Close($)	AMD Adj· Close($)	ORCL Adj· Close($)	AAPL Return	AMD Return	ORCL Return
1	4.16	17.57	4.32	0.0962	−0.5537	−0.1074
2	4.58	10.1	3.88	0.8104	0.1272	0.8666
3	10.3	11.47	9.23	0.9236	0.4506	0.9956
4	25.94	18	24.98	−0.8753	0.3124	0.1533
5	10.81	24.6	29.12	0.1340	−0.4270	−0.5230
6	12.36	16.05	17.26	−0.5432	−1.1194	−0.3610
7	7.18	5.24	12.03	0.4517	1.0424	0.1416
8	11.28	14.86	13.86	1.2263	0.0613	−0.0065
9	38.45	15.8	13.77	0.6749	0.9729	−0.0912
10	75.51	41.8	12.57			

Data Source: CSI

Web site: www.csidata.com

24. **Markowitz Portfolio Optimization.** Formulate and solve the Markowitz portfolio optimization model to minimize portfolio variance subject to a required expected return of 10% that was defined in equations (16.8) through (16.17) using the return data from Problem 23. In this case, the nine scenarios correspond to the yearly returns from Years 1 through 9. Treat each scenario as being equally likely, and use the scenario returns that were calculated in Problem 23. **LO 1, 2, 4**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file*
stockreturnsindex

25. **Creating a Stock Index Fund.** The file *stockreturnsindex* contains nine years of returns for Apple (AAPL), American Medical Devices (AMD), Oracle Corporation (ORCL), and the Information Technology S&P Index. Construct a portfolio from AAPL, AMD, and ORCL that matches the return of the Information Technology S&P index as closely as possible. *Hint:* Minimize the sum of the squared differences between the portfolio return and the return of the Information Technology S&P Index. **LO 1, 2**

26. **Cable Bundle Pricing.** Queen City Cable (QCC) is interested in trying bundle pricing for its offerings of cable television, cell phone service, and internet service, with the goal of maximizing revenue. The following table shows the current proposal from QCC pricing analysts for each type of service per month.

Individual Services		Two-Service Bundles		Three-Service Bundle	
Cable TV	$90	Cable TV + Cell Phone	$225	Cable TV + Cell Phone + Internet	$300
Cell Phone	$150	Cable TV + Internet	$225		
Internet	$80	Cell Phone + Internet	$225		

The analysts feel that no single service's monthly price should be more than $100 for cable TV, no more than $200 for cell phone service, and no more than $100 for internet service. Also, no two-service bundle's monthly price should be more than $250, and the bundle price for all three services per month should not be more than $350.

DATA *file*
cablebundle

The file *cablebundle* contains data from 100 survey respondents regarding what they would be willing to pay individually per month for cable TV, cellphone service, and internet service. **LO 7**

MODEL *file*
mtairy

a. Build a spreadsheet model similar to the model in the file *mtairy* that was discussed in section 16.6 to maximize revenue.
b. Solve the model you constructed in part a using the Evolutionary method with the default options but with the Random Seed to 1989. If Solver asks if you want to stop or continue, select **Stop** and accept the current solution. What is the maximum revenue?
c. What is the percentage improvement in revenue from the solution found using the Evolutionary method over the proposed pricing strategy?

27. **Asymmetric Traveling Salesperson Problem.** A well-studied optimization problem is the traveling salesperson problem (TSP). Given travel distances between N cities, the TSP problem is to construct a tour that has minimal total distance traveled. A tour is defined as a loop that contains all N cities in which you enter and leave each city once, and return home to the city in which you began. The asymmetric traveling salesperson problem refers to the data not being symmetric, that is, the distance from city A to city B is not necessarily the same as the distance from city B to city A. The TSP problem is difficult to solve to optimality because integer-programming-based methods result in subtours forming rather than one large tour. However, a heuristic approach using the INDEX function in Excel with the Evolutionary method and a special constraint called "All Different" can yield good solutions. The "All Different" constraint ensures that the decision variables each take on a different integer value.

Integer programming models are discussed in Chapter 15.

The file *atsp* contains a matrix with the distances in miles between 15 cities and a model for solving the associated TSP:

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	**Traveling Salesperson Problem**																
2																	
3	**Data**																
4		To															
5	From	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
6	1	0	2590	745	3301	3887	554	2501	1364	1932	2804	2112	2954	2868	774	1891	
7	2	966	0	2443	2961	4078	1178	1960	1767	2119	1943	1089	3592	506	2080	2464	
8	3	4661	1080	0	629	3978	4091	4009	3763	1000	2219	790	793	3676	951	3490	
9	4	4145	1547	5000	0	3523	1595	990	959	3889	3667	2693	2118	4893	4201	1756	
10	5	4681	2319	2477	1375	0	4618	2949	2100	4828	2032	1080	3042	1654	3317	684	
11	6	4568	3813	4536	3859	3725	0	1341	3162	2768	2759	3105	3344	4046	3518	1006	
12	7	2480	3686	564	1567	2984	2587	0	2134	1019	1823	1476	3950	902	2621	2465	
13	8	2218	4233	1789	674	639	2621	788	0	940	1692	4460	3262	2137	3475	4800	
14	9	728	561	3819	1282	3241	2088	2359	4981	0	2516	1578	1426	3313	3560	3046	
15	10	4167	3693	3204	4308	4302	1080	1250	4992	4595	0	4850	540	2616	3865	4380	
16	11	2214	3972	4589	3461	723	4350	4732	2331	2001	3053	0	4665	958	4035	507	
17	12	504	3149	3796	4465	1452	4623	4363	1272	4062	4113	3998	0	3332	1476	2660	
18	13	2271	3255	505	970	2659	1984	3720	2501	1621	981	4754	3758	0	1434	756	
19	14	4910	3055	2290	2070	3996	3941	1439	2957	1498	675	1288	1560	4563	0	2071	
20	15	1854	1613	4067	4646	2075	1678	2308	4995	3953	860	4686	2999	4281	1233	0	
21																	
22	**Model**																
23																	
24	Total Distance Traveled	38,940															
25																	
26	Tour Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1
27	City	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1
28			2590	2443	629	3523	4618	1341	2134	940	2516	4850	4665	3332	1434	2071	1854
29																	

The current solution in row 27 is to simply travel the cities in the order they are numbered. As shown in cell B24, the total distance traveled for this tour is 38,940 miles. For any proposed tour (ordering of the cities in row 27, we use the INDEX function to get the distance between pairs of cities in the tour (note in the following figure we have hidden columns D through P).

	A	B	C	Q
22	**Model**			
23				
24	Total Distance Traveled	=SUM(C28:Q28)		
25				
26	Tour Position	1	2	1
27	City	1	2	=B27
28			=INDEX($B$6:$P$20,B27,C27)	=INDEX($B$6:$P$20,P27,Q27)
29				

MODEL file
atsp

The INDEX function works as follows: =INDEX($B$6:$P$20,B27,C27) finds the element in the matrix defined by B6:P20 in the row defined by the value in cell B27 and the column defined by the value cell C27. For example, in cell C28 the value returned is the element in the matrix in the first row and the second column, which is 2590. In cell Q27, the formula =B27 forces the tour to end where it started and the formula in cell B24, =SUM(C28:Q28) adds up the distances to get the total tour length. **LO 7**

a. Use the Evolutionary method to solve this problem. Minimize B24 with Changing Variable Cells B27:P27, and an All Different constraint for cells B27:P27 (add a constraint and select "dif" as the constraint type). Keep the default options, but set the Random Seed to 15. Give the best tour found and its total distance.
b. What percentage improvement is gained by the solution found versus the original tour which followed the cities in order of their number?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

28. Symmetric Traveling Salesperson Problem. Refer to Problem 27 for a definition of the asymmetric traveling salesperson problem. The *symmetric* traveling salesperson problem is the same problem, but with symmetric distances. That is, the distance from city A to city B is the same as the distance from city B to city A. The file *stsp* contains an upper triangular matrix of distance between pairs of 10 cities as follows. **LO 7**

	To									
From	1	2	3	4	5	6	7	8	9	10
1	0	270	198	302	336	489	328	347	482	490
2		0	221	332	488	281	287	406	396	388
3			0	429	189	215	190	274	297	354
4				0	466	137	318	297	398	448
5					0	160	237	463	204	314
6						0	189	357	155	195
7							0	167	126	464
8								0	439	139
9									0	452
10										0

MODEL *file*
atsp

DATA *file*
stsp

a. Modify the modeling approach from problem 26 in the file *astp* to build a model that can be used to solve the symmetric TSP for the data in file *stsp*. *Hint:* the distance from city 4 to 2 can be obtained from the distance given form city 2 to city 4 (332).
b. What is the total distance traveled for the tour that follows the cities in the order they are numbered?
c. Use the Evolutionary method in Excel Solver to improve the solution used in part b. Use all default options except the Random Seed, which you should set to 58. What is the tour length found?
d. What is the percentage improvement due to using the Evolutionary method over the solution used in part b?

Case Problem: Portfolio Optimization with Transaction Costs

Hauck Financial Services has a number of passive, buy-and-hold clients. For these clients, Hauck offers an investment account whereby clients agree to put their money into a portfolio of mutual funds that is rebalanced once a year. When the rebalancing occurs, Hauck determines the mix of mutual funds in each investor's portfolio by solving an extension of the Markowitz portfolio model that incorporates transaction costs. Investors are charged a small transaction cost for the annual rebalancing of their portfolio. For simplicity, assume the following:

- At the beginning of the time period (in this case one year), the portfolio is rebalanced by buying and selling Hauck mutual funds.
- The transaction costs associated with buying and selling mutual funds are paid at the beginning of the period when the portfolio is rebalanced, which, in effect, reduces the amount of money available to reinvest.
- No further transactions are made until the end of the time period, at which point the new value of the portfolio is observed.
- The transaction cost is a linear function of the dollar amount of mutual funds bought or sold.

Jean Delgado is one of Hauck's buy-and-hold clients. We briefly describe the model as it is used by Hauck for rebalancing her portfolio. The mix of mutual funds that are being considered for her portfolio are a foreign stock fund (*FS*), an intermediate-term bond fund (*IB*), a large-cap growth fund (*LG*), a large-cap value fund (*LV*), a small-cap growth fund (*SG*), and a small-cap value fund (*SV*). In the traditional Markowitz model, the variables are usually interpreted as the proportion of the portfolio invested in the asset represented by the variable. For example, *FS* is the proportion of the portfolio invested in the foreign stock fund. However, it is equally correct to interpret *FS* as the dollar amount invested in the foreign stock fund. Then $FS = 25{,}000$ implies that \$25,000 is invested in the foreign stock fund. Based on these

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

assumptions, the initial portfolio value must equal the amount of money spent on transaction costs plus the amount invested in all the assets after rebalancing; that is,

Initial portfolio value = amount invested in all assets after rebalancing + transaction costs

The extension of the Markowitz model that Hauck uses for rebalancing portfolios requires a balance constraint for each mutual fund. This balance constraint is

Amount invested in fund i = initial holding of fund i + amount of fund i purchased − amount of fund i sold

DATA*file* stockprices

Using this balance constraint requires three additional variables for each fund: one for the amount invested prior to rebalancing, one for the amount sold, and one for the amount purchased. For instance, the balance constraint for the foreign stock fund is

$$FS = FS_START + FS_BUY - FS_SELL$$

Jean Delgado has $100,000 in her account prior to the annual rebalancing, and she has specified a minimum acceptable return of 10%. Hauck plans to use the following model to rebalance Jean's portfolio. The complete model with transaction costs is

$$\text{Min } \tfrac{1}{5}\sum_{s=1}^{5}(R_s - \bar{R})^2$$

s.t.

$$\begin{aligned}
0.1006FS + 0.1764IB + 0.3241LG + 0.3236LV + 0.3344SG + 0.2456SV &= R_1\\
0.1312FS + 3.2500IB + 0.1871LG + 0.2061LV + 0.1940SG + 0.2532SV &= R_2\\
0.1347FS + 0.0751IB + 0.3328LG + 0.1293LV + 0.3850SG - 0.0670SV &= R_3\\
0.4542FS - 0.0133IB + 0.4146LG + 0.0706LV + 0.5868SG + 0.0543SV &= R_4\\
-0.2193FS + 0.0736IB - 0.2326LG - 0.0537LV - 0.0902SG + 0.1731SV &= R_5\\
\tfrac{1}{5}\sum_{s=1}^{5} R_s &= \bar{R}\\
\bar{R} &\geq 10{,}000\\
FS + IB + LG + LV + SG + SV + TRANS_COST &= 100{,}000\\
FS_START + FS_BUY - FS_SELL &= FS\\
IB_START + IB_BUY - IB_SELL &= IB\\
LG_START + LG_BUY - LG_SELL &= LG\\
LV_START + LV_BUY - LV_SELL &= LV\\
SG_START + SG_BUY - SG_SELL &= SG\\
SV_START + SV_BUY - SV_SELL &= SV\\
TRANS_FEE * (FS_BUY + FS_SELL + IB_BUY + IB_SELL + & \\
LG_BUY + LG_SELL + LV_BUY + LV_SELL + SG_BUY + & \\
SG_SELL + SV_BUY + SV_SELL) &= TRANS_COST\\
FS_START &= 10{,}000\\
IB_START &= 10{,}000\\
LG_START &= 10{,}000\\
LV_START &= 40{,}000\\
SG_START &= 10{,}000\\
SV_START &= 20{,}000\\
TRANS_FEE &= 0.01\\
FS, IB, LG, LV, SG, SV &\geq 0
\end{aligned}$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notice that the transaction fee is set at 1% in the model (the last constraint) and that the transaction cost for buying and selling shares of the mutual funds is a linear function of the amount bought and sold. With this model, the transaction costs are deducted from the client's account at the time of rebalancing and thus reduce the amount of money invested. The solution for Jean's rebalancing problem is shown as part of the Managerial Report. **LO 1, 2, 4**

Managerial Report

Assume that you are a financial analytics specialist newly hired by Hauck Financial Services. One of your first tasks is to review the portfolio rebalancing model in order to resolve a dispute with Jean Delgado. Jean has had one of the Hauck passively managed portfolios for the past five years and has complained that she is not getting the rate of return of 10% that she specified. After reviewing annual statements for the past five years, she feels that she is actually getting less than 10% on average.

1. According to the following Model Solution, $IB_BUY = \$41{,}268.51$. How much in transaction costs did Jean pay for purchasing additional shares of the intermediate-term bond fund?
2. Based on the Model Solution, what is the total transaction cost associated with rebalancing Jean's portfolio?
3. After paying transactions costs, how much did Jean have invested in mutual funds after her portfolio was rebalanced?
4. According to the Model Solution, $IB = \$51{,}268.51$. How much can Jean expect to have in the intermediate-term bond fund at the end of the year?
5. According to the Model Solution, the expected return of the portfolio is $10,000. What is the expected dollar amount in Jean's portfolio at the end of the year? Can she expect to earn 10% on the $100,000 she had at the beginning of the year?

MODEL SOLUTION

Optimal	Objective Value		
	27219457.356		
Variable	**Value**	**Variable**	**Value**
R_1	18,953.280	*IB_START*	10,000.000
$\bar{R}$	10,000.000	*IB_BUY*	41,268.510
R_2	11,569.210	*IB_SELL*	0.000
R_3	5,663.961	*LG_START*	10,000.000
R_4	9,693.921	*LG_BUY*	0.000
R_5	4,119.631	*LG_SELL*	5,060.688
FS	15,026.860	*LV_START*	40,000.000
IB	51,268.510	*LV_BUY*	0.000
LG	4,939.312	*LV_SELL*	40,000.000
LV	0.000	*SG_START*	10,000.000
SG	0.000	*SG_BUY*	0.000
SV	27,675.000	*SG_SELL*	10,000.000
TRANS_COST	1,090.311	*SV_START*	20,000.000
FS_START	10,000.000	*SV_BUY*	7,675.004
FS_BUY	5,026.863	*SV_SELL*	0.000
FS_SELL	0.000	*TRANS_FEE*	0.010

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

6. It is now time to prepare a report to management to explain why Jean did not earn 10% each year on her investment. Make a recommendation in terms of a revised portfolio model that can be used so that Jean can have an expected portfolio balance of $110,000 at the end of next year. Prepare a report that includes a modified optimization model that will give an expected return of 10% on the amount of money available at the beginning of the year before paying the transaction costs. Explain why the current model does not do this.
7. Solve the formulation in part (6) for Jean. How does the portfolio composition differ from that of the Model Solution?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Chapter 17

Decision Analysis

Contents

Analytics In Action:
U.S. Centers for Disease Control and Prevention (CDC)

17.1 Problem Formulation
- Payoff Tables
- Decision Trees

17.2 Decision Analysis Without Probabilities
- Optimistic Approach
- Conservative Approach
- Minimax Regret Approach

17.3 Decision Analysis with Probabilities
- Expected Value Approach
- Risk Analysis
- Sensitivity Analysis

17.4 Decision Analysis with Sample Information
- Expected Value of Sample Information
- Expected Value of Perfect Information

17.5 Computing Branch Probabilities with Bayes' Theorem

17.6 Utility Theory
- Utility and Decision Analysis
- Utility Functions
- Exponential Utility Function

Summary 923
Glossary 923
Problems 925

Learning Objectives

After completing this chapter, you will be able to:

LO 1 Construct and interpret payoff tables and decision trees for a decision problem.

LO 2 Use the optimistic, conservative, and minimax regret approaches to determine the optimal decision when probabilities for the states of nature are unknown.

LO 3 Use the expected value approach to determine the optimal decision when probabilities for the states of nature are known.

LO 4 Construct a risk profile of a decision problem when probabilities for the states of nature are known.

LO 5 Use sensitivity analysis to determine the range of probabilities for which each decision alternative has the largest expected value.

LO 6 Calculate and interpret the expected value of sample information and the expected value of perfect information for a decision problem.

LO 7 Use Bayes' theorem to calculate and interpret posterior probabilities for a decision problem.

LO 8 Calculate the utility of payoffs in a decision problem and use the expected utility approach to determine the optimal decision.

LO 9 Graph the utility function for a decision maker and classify the decision maker as a risk avoider, a risk taker, or risk neutral.

LO 10 Use the exponential utility function to represent a decision maker's utility and use it to determine the optimal decision.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Analytics in Action

U.S. Centers for Disease Control and Prevention (CDC)*

Over the course of history, polio has been one of the most terrifying diseases known to humans. Polio is caused by the poliovirus. Wild polioviruses (WPVs) refer to polioviruses that exist in the wild, rather than viruses that are derived from vaccines. While cases of infection due to vaccine-derived polioviruses have occasionally been reported in the United States, including a case in New York State in 2022, no cases of polio due to WPVs have been reported in the United States since 1979. WPVs are generally stronger strains of the virus which can lead to serious long-term effects, including paralysis.

WPVs are typically transmitted when contaminated water or food is ingested by an uninfected person, which can be common in areas with poor hygiene or poor sanitation. In the 1950s, polio vaccines were developed separately by Jonas Salk and Albert Sabin. These vaccines led to the eradication of WPV transmission in the United States by 1979 and throughout the western hemisphere by 1991. However, polio remains endemic to several countries around the world, which puts all other populations at risk of infection.

The Global Polio Eradication Initiative (GPEI) seeks to eradicate polio worldwide. It is financed by a variety of public and private donors including the World Health Organization, the U.S. Centers for Disease Control and Prevention (CDC), and the Bill & Melinda Gates Foundation. The CDC collaborated with Kid Risk, Inc. to use analytics tools to evaluate the policy options available for pursuing global polio eradication in terms of benefits, risks, and costs. The researchers used a variety of decision analysis tools, including decision trees, to consider the many different options available to potentially eradicate polio around the world. These policy decisions include such options as using oral poliovirus vaccines (OPVs), inactivated poliovirus vaccines (IPVs), and whether to continue routine vaccinations after WPV transmission has been eradicated. The decision trees allowed the researchers to capture much of the complexity inherent in such a large-scale decision problem and to evaluate the relative risks, benefits, and costs of different policy options. The uncertainties in costs were modeled using 47 different probability distributions for cost inputs, and cost data were collected through extensive data collection efforts led by the Global Polio Laboratory Network.

The findings from the decision analysis model led researchers to understand that the speed with which medical professionals respond to a polio outbreak can be more effective than having extremely high rates of WPV immunity through vaccinations. The models also allowed the researchers to calculate the cost tradeoffs of eradicating polio worldwide versus responding to greater polio outbreaks. The research provided extensive evidence of the value of the GPEI by showing that net benefits of countries covered by the GPEI approach $40–50 billion. The work done by these researchers was recognized with the 2014 INFORMS Franz Edelman Award based on the enormous financial and global health benefits derived from these analytical models.

*Based on K. M. Thompson, R. J. Duintjer Tebbens, M. A. Pallansch, S. G. F. Wassilak, and S. L. Cochi, "Polio eradicators use integrated analytical models to make better decisions," *Interfaces*, 45, no. 1 (January–February 2015): 1–16.

Ultimately, business analytics is about making better decisions. The tools and techniques we have introduced previously are designed to aid a decision maker in analyzing existing data, predicting future behavior, and recommending decisions. This chapter introduces a field known as decision analysis that can be used to develop an optimal strategy when a decision maker is faced with several decision alternatives and an uncertain or risk-filled pattern of future events.

Decision analysis techniques are used widely in many different settings. The Analytics in Action for this chapter discusses how the U.S. Centers for Disease Control and Prevention used decision analysis to evaluate options for trying to eradicate polio worldwide. Other federal agencies in the United States have used decision analysis to evaluate the potential risks from terrorist attacks and to recommend counterterrorism strategies. The State of North Carolina used decision analysis in evaluating whether to implement a medical screening test to detect metabolic disorders in newborns. Decision analysis is used by many business organizations, such as Procter & Gamble, to analyze new product introductions, marketing decisions, and much more. For example, by evaluating the different naming options and understanding the potential sources of uncertainty, Procter & Gamble used decision analysis techniques to help choose the best brand name when they introduced Crest White Strips.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Even when a careful decision analysis has been conducted, uncertainty about future events means that the final outcome is not completely under the control of the decision maker. In some cases, the selected decision alternative may provide good or excellent results. In other cases, a relatively unlikely future event may occur, causing the selected decision alternative to provide only fair or even poor results. The risk associated with any decision alternative is a direct result of the uncertainty associated with the final outcome. A good decision analysis includes careful consideration of risk. Through risk analysis, the decision maker is provided with probability information about the favorable as well as the unfavorable outcomes that may occur.

We begin the study of decision analysis by considering problems that involve reasonably few decision alternatives and reasonably few possible future events. Payoff tables and decision trees are introduced to provide a structure for the decision problem and to illustrate the fundamentals of decision analysis. Decision trees are used to analyze more complex problems and to identify an optimal sequence of decisions, referred to as an *optimal decision strategy*. Sensitivity analysis shows how changes in various aspects of the problem affect the recommended decision alternative. We return to the use of Bayes' theorem for calculating the probabilities of future events and incorporating additional information about the decisions. We conclude this chapter with a discussion of utility and decision analysis that expands on different attitudes toward risk taken by decision makers.

Bayes' theorem is introduced in Chapter 5.

17.1 Problem Formulation

The first step in the decision analysis process is problem formulation. We begin with a verbal statement of the problem. We then identify the **decision alternatives**; the uncertain future events, referred to as **chance events**; and the **outcomes** associated with each combination of decision alternative and chance event outcome. Let us begin by considering a construction project of the Pittsburgh Development Corporation (PDC).

PDC purchased land that will be the site of a new luxury condominium complex. The location provides a spectacular view of downtown Pittsburgh and the Golden Triangle, where the Allegheny and Monongahela Rivers meet to form the Ohio River. PDC plans to price the individual condominium units between $300,000 and $1,400,000.

PDC commissioned preliminary architectural drawings for three different projects: one with 30 condominiums, one with 60 condominiums, and one with 90 condominiums. The financial success of the project depends on the size of the condominium complex and the chance event concerning the demand for the condominiums. The statement of the PDC decision problem is to select the size of the new luxury condominium project that will lead to the largest profit given the uncertainty concerning the demand for the condominiums.

Given the statement of the problem, it is clear that the decision is to select the best size for the condominium complex. PDC has the following three decision alternatives:

d_1 = a small complex with 30 condominiums
d_2 = a medium complex with 60 condominiums
d_3 = a large complex with 90 condominiums

A factor in selecting the best decision alternative is the uncertainty associated with the chance event concerning the demand for the condominiums. When asked about the possible demand for the condominiums, PDC's president acknowledged a wide range of possibilities but decided that it would be adequate to consider two possible chance event outcomes: a strong demand and a weak demand.

In decision analysis, the possible outcomes for a chance event are referred to as the **states of nature**. The states of nature are defined so that they are mutually exclusive (no more than one can occur) and collectively exhaustive (at least one must occur); thus, one and only one of the possible states of nature will occur. For the PDC problem, the chance event concerning the demand for the condominiums has two states of nature:

s_1 = strong demand for the condominiums
s_2 = weak demand for the condominiums

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 17.1 Payoff Table for the PDC Condominium Project ($ millions)

Decision Alternative	State of Nature: Strong Demand, s_1	State of Nature: Weak Demand, s_2
Small complex, d_1	8	7
Medium complex, d_2	14	5
Large complex, d_3	20	−9

Management must first select a decision alternative (complex size); then a state of nature follows (demand for the condominiums), and finally an outcome will occur. In this case, the outcome is PDC's profit.

Payoff Tables

Payoffs can be expressed in terms of profit, cost, time, distance, or any other measure appropriate for the decision problem being analyzed.

Given the three decision alternatives and the two states of nature, which complex size should PDC choose? To answer this question, PDC will need to know the outcome associated with each possible combination of decision alternative and state of nature. In decision analysis, we refer to the outcome resulting from a specific combination of a decision alternative and a state of nature as a **payoff**. A table showing payoffs for all combinations of decision alternatives and states of nature is a **payoff table**.

Because PDC wants to select the complex size that provides the largest profit, profit is used as the outcome. The payoff table with profits (in millions of dollars) is shown in Table 17.1. Note, for example, that if a medium complex is built and demand turns out to be strong, a profit of $14 million will be realized. We will use the notation V_{ij} to denote the payoff associated with decision alternative i and state of nature j. Using Table 17.1, $V_{31} = 20$ indicates that a payoff of $20 million occurs if the decision is to build a large complex (d_3) and the strong demand state of nature (s_1) occurs. Similarly, $V_{32} = -9$ indicates a loss of $9 million if the decision is to build a large complex (d_3) and the weak demand state of nature (s_2) occurs.

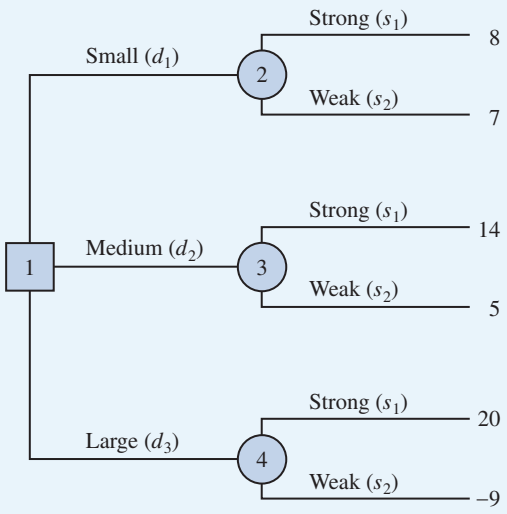
Decision Trees

A **decision tree** provides a graphical representation of the decision-making process. Figure 17.1 presents a decision tree for the PDC problem. Note that the decision tree shows the natural or logical progression that will occur over time. First, PDC must make a decision regarding the size of the condominium complex (d_1, d_2, or d_3). Then, after the decision is implemented, either state of nature s_1 or s_2 will occur. The number at each end point of the tree indicates that the payoff is associated with a particular sequence. For example, the topmost payoff of 8 indicates that an $8 million profit is anticipated if PDC constructs a small condominium complex (d_1) and demand turns out to be strong (s_1). The next payoff of 7 indicates an anticipated profit of $7 million if PDC constructs a small condominium complex (d_1) and demand turns out to be weak (s_2). Thus, the decision tree provides a graphical depiction of the sequences of decision alternatives and states of nature that provide the six possible payoffs for PDC.

The decision tree in Figure 17.1 shows four nodes, numbered 1 to 4. **Nodes** are used to represent decisions and chance events. Squares are used to depict **decision nodes**, circles are used to depict **chance nodes**. Thus, node 1 is a decision node, and nodes 2, 3, and 4 are chance nodes. The **branches** connect the nodes; those leaving the decision node correspond to the decision alternatives. The branches leaving each chance node correspond to the states of nature. The outcomes (payoffs) are shown at the end of the states-of-nature branches. We now turn to the question: How can the decision maker use the information in the payoff table or the decision tree to select the best decision alternative? Several approaches may be used and are covered in the remaining sections of this chapter.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 17.1 Decision Tree for the PDC Condominium Project ($ millions)



Notes + Comments

1. The first step in solving a complex problem is to decompose the problem into a series of smaller subproblems. Decision trees provide a useful way to decompose a problem and illustrate the sequential nature of the decision process.
2. People often view the same problem from different perspectives. Thus, the discussion regarding the development of a decision tree may provide additional insight into the problem.

17.2 Decision Analysis Without Probabilities

In this section, we consider approaches to decision analysis that do not require knowledge of the probabilities of the states of nature. These approaches are appropriate in situations in which a simple best-case and worst-case analysis is sufficient and in which the decision maker has little confidence in their ability to assess the probabilities. Because different approaches sometimes lead to different decision recommendations, the decision maker must understand the approaches available and then select the specific approach that, according to the judgment of the decision maker, is the most appropriate.

Optimistic Approach

For a maximization problem, the optimistic approach often is referred to as the maximax approach; for a minimization problem, the corresponding terminology is minimin.

The **optimistic approach** evaluates each decision alternative in terms of the *best* payoff that can occur. The decision alternative that is recommended is the one that provides the best possible payoff. For a problem in which maximum profit is desired, as in the PDC problem, the optimistic approach would lead the decision maker to choose the alternative corresponding to the largest profit. For problems involving minimization, this approach leads to choosing the alternative with the smallest payoff.

To illustrate the optimistic approach, we use it to develop a recommendation for the PDC problem. First, we determine the maximum payoff for each decision alternative; then we select the decision alternative that provides the overall maximum payoff. These steps

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 17.2 Maximum Payoff for Each PDC Decision Alternative

Decision Alternative	Maximum Payoff ($ millions)	
Small complex, d_1	8	
Medium complex, d_2	14	
Large complex, d_3	20	← Maximum of the maximum payoff values

systematically identify the decision alternative that provides the largest possible profit. Table 17.2 illustrates these steps.

Because 20, corresponding to d_3, is the largest payoff, the decision to construct the large condominium complex is the recommended decision alternative using the optimistic approach.

Conservative Approach

For a maximization problem, the conservative approach often is referred to as the maximin approach; for a minimization problem the corresponding terminology is minimax.

The **conservative approach** evaluates each decision alternative in terms of the *worst* payoff that can occur. The decision alternative recommended is the one that provides the best of the worst possible payoffs. For a problem in which the output measure is profit, as in the PDC problem, the conservative approach would lead the decision maker to choose the alternative that maximizes the minimum possible profit that could be obtained. For problems involving minimization (e.g., when the output measure is cost instead of profit), this approach identifies the alternative that will minimize the maximum payoff.

To illustrate the conservative approach, we use it to develop a recommendation for the PDC problem. First, we identify the minimum payoff for each of the decision alternatives; then we select the decision alternative that maximizes the minimum payoff. Table 17.3 illustrates these steps for the PDC problem.

Because 7, corresponding to d_1, yields the maximum of the minimum payoffs, the decision alternative of a small condominium complex is recommended. This decision approach is considered conservative because it identifies the worst possible payoffs and then recommends the decision alternative that avoids the possibility of extremely "bad" payoffs. In the conservative approach, PDC is guaranteed a profit of at least $7 million. Although PDC may make more, it *cannot* make less than $7 million.

Minimax Regret Approach

In decision analysis, **regret** is the difference between the payoff associated with a *particular* decision alternative and the payoff associated with the decision that would yield the most desirable payoff for a given state of nature. Thus, regret represents how much potential payoff one would forgo by selecting a *particular* decision alternative, given that a specific state of nature will occur. This is why regret is often referred to as **opportunity loss**.

As its name implies, under the **minimax regret approach** to decision analysis, one would choose the decision alternative that minimizes the maximum state of regret that could occur over all possible states of nature. This approach is neither purely optimistic nor

Table 17.3 Minimum Payoff for Each PDC Decision Alternative

Decision Alternative	Minimum Payoff ($ millions)	
Small complex, d_1	7	← Maximum of the minimum payoff values
Medium complex, d_2	5	
Large complex, d_3	−9	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

purely conservative. Let us illustrate the minimax regret approach by showing how it can be used to select a decision alternative for the PDC problem.

Suppose that PDC constructs a small condominium complex (d_1) and demand turns out to be strong (s_1). Table 17.1 showed that the resulting profit for PDC would be \$8 million. However, given that the strong demand state of nature (s_1) has occurred, we realize that the decision to construct a large condominium complex (d_3), yielding a profit of \$20 million, would have been the best decision. The difference between the payoff for the best decision alternative (\$20 million) and the payoff for the decision to construct a small condominium complex (\$8 million) is the regret or opportunity loss associated with decision alternative d_1 when state of nature s_1 occurs; thus, for this case, the opportunity loss or regret is \$20 million − \$8 million = \$12 million. Similarly, if PDC makes the decision to construct a medium condominium complex (d_2) and the strong demand state of nature (s_1) occurs, the opportunity loss, or regret, associated with d_2 would be \$20 million − \$14 million = \$6 million. Of course, if PDC chooses to construct a large complex (d_3) and demand is strong, they would have no regret.

In general, the following expression represents the opportunity loss, or regret:

Regret (Opportunity Loss)

$$R_{ij} = \left|V_j^* - V_{ij}\right| \tag{17.1}$$

where

R_{ij} = the regret associated with decision alternative d_i and state of nature s_j

V_j^* = the payoff value corresponding to the best decision for the state of nature s_j

V_{ij} = the payoff corresponding to decision alternative d_i and state of nature s_j

Note the role of the absolute value in equation (17.1). For minimization problems, the best payoff, V_j^*, is the smallest entry in column j. Because this value always is less than or equal to V_{ij}, the absolute value of the difference between V_j^* and V_{ij} ensures that the regret is always the magnitude of the difference.

Using equation (17.1) and the payoffs in Table 17.1, we can compute the regret associated with each combination of decision alternative d_i and state of nature s_j. Because the PDC problem is a maximization problem, V_j^* will be the largest entry in column j of the payoff table. Thus, to compute the regret, we simply subtract each entry in a column from the largest entry in the column. Table 17.4 shows the opportunity loss, or regret, table for the PDC problem.

The next step in applying the minimax regret approach is to list the maximum regret for each decision alternative; Table 17.5 shows the results for the PDC problem. Selecting the decision alternative with the *minimum* of the *maximum* regret values—hence, the name *minimax regret*—yields the minimax regret decision. For the PDC problem, the alternative to construct the medium condominium complex, with a corresponding maximum regret of \$6 million, is the recommended minimax regret decision.

Table 17.4 Opportunity Loss, or Regret, Table for the PDC Condominium Project (\$ millions)

	State of Nature	
Decision Alternative	**Strong Demand, s_1**	**Weak Demand, s_2**
Small complex, d_1	12	0
Medium complex, d_2	6	2
Large complex, d_3	0	16

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 17.5 Maximum Regret for Each PDC Decision Alternative

Decision Alternative	Maximum Regret ($ millions)	
Small complex, d_1	12	
Medium complex, d_2	6	← Minimum of the maximum regret
Large complex, d_3	16	

Note that the three approaches discussed in this section provide different recommendations, which in itself is not bad. It simply reflects the difference in decision-making philosophies that underlie the various approaches. Ultimately, the decision maker will have to choose the most appropriate approach and then make the final decision accordingly. The main criticism of the approaches discussed in this section is that they do not consider any information about the probabilities of the various states of nature. In the next section, we discuss an approach that utilizes probability information in selecting a decision alternative.

17.3 Decision Analysis with Probabilities

Expected Value Approach

In many decision-making situations, we can obtain probability assessments for the states of nature. When such probabilities are available, we can use the **expected value approach** to identify the best decision alternative. Let us first define the expected value of a decision alternative and then apply it to the PDC problem.

Let

$$N = \text{the number of states of nature}$$
$$P(s_j) = \text{the probability of state of nature } s_j$$

Because one and only one of the N states of nature can occur, the probabilities must satisfy two conditions:

$$P(s_j) \geq 0 \quad \text{for all states of nature}$$
$$\sum_{j=1}^{N} P(s_j) = P(s_1) + P(s_2) + \cdots + P(s_N) = 1$$

The **expected value (EV)** of decision alternative d_i is defined as follows:

Expected Value of Decision Alternative d_i

$$\text{EV}(d_i) = \sum_{j=1}^{N} P(s_j) V_{ij} \tag{17.2}$$

In words, the expected value of a decision alternative is the sum of weighted payoffs for the decision alternative. The weight for a payoff is the probability of the associated state of nature and therefore the probability that the payoff will occur. Let us return to the PDC problem to see how the expected value approach can be applied.

PDC is optimistic about the potential for the luxury high-rise condominium complex. Suppose that this optimism leads to an initial subjective probability assessment of 0.8 that demand will be strong (s_1) and a corresponding probability of 0.2 that demand will be weak (s_2). Thus, $P(s_1) = 0.8$ and $P(s_2) = 0.2$. Using the payoff values in Table 17.1 and equation (17.2), we compute the expected value for each of the three decision alternatives as follows:

$$\text{EV}(d_1) = 0.8(8) + 0.2(7) = 7.8$$
$$\text{EV}(d_2) = 0.8(14) + 0.2(5) = 12.2$$
$$\text{EV}(d_3) = 0.8(20) + 0.2(-9) = 14.2$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Thus, using the expected value approach, we find that the large condominium complex, with an expected value of $14.2 million, is the recommended decision.

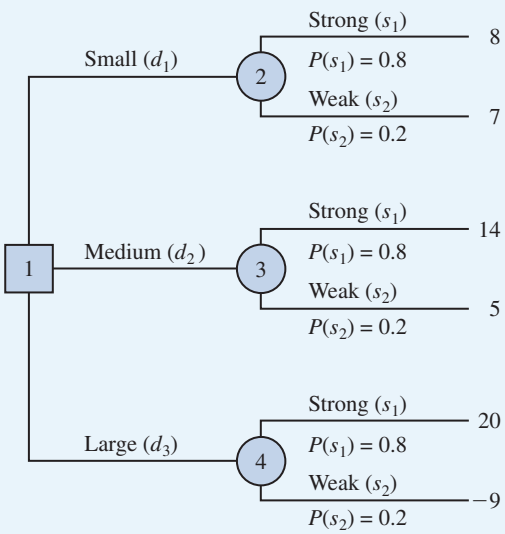
The calculations required to identify the decision alternative with the best expected value can be conveniently carried out on a decision tree. Figure 17.2 shows the decision tree for the PDC problem with state-of-nature branch probabilities. Working backward through the decision tree, we first compute the expected value at each chance node. In other words, at each chance node, we weight each possible payoff by its probability of occurrence. By doing so, we obtain the expected values for nodes 2, 3, and 4, as shown in Figure 17.3.

Because the decision maker controls the branch leaving decision node 1 and because we are trying to maximize the expected profit, the best decision alternative at node 1 is d_3. Thus, the decision tree analysis leads to a recommendation of d_3, with an expected value of $14.2 million. Note that this recommendation is also obtained with the expected value approach in conjunction with the payoff table.

Other decision problems may be substantially more complex than the PDC problem, but if a reasonable number of decision alternatives and states of nature are present, you can use the decision tree approach outlined here. First, draw a decision tree consisting of decision nodes, chance nodes, and branches that describe the sequential nature of the problem. If you use the expected value approach, the next step is to determine the probabilities for each of the states of nature and compute the expected value at each chance node. Then select the decision branch leading to the chance node with the best expected value. The decision alternative associated with this branch is the recommended decision.

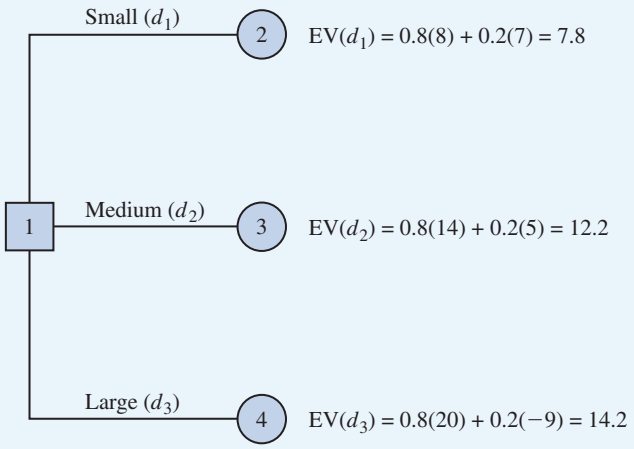
In practice, obtaining precise estimates of the probabilities for each state of nature is often impossible. In some cases, where similar decisions have been made many times in the past, one may use historical data to estimate the probabilities for the different states of nature. However, often there are little, or no, historical data to guide the estimates of these probabilities. In these cases, we may have to rely on subjective estimates to determine the probabilities for the states of nature. When relying on subjective estimates, we often want to get more than one estimate because many studies have shown that even knowledgeable experts are often overly optimistic in their estimates. It is also particularly important when dealing with subjective probability estimates to perform risk analysis and sensitivity analysis, as we will explain.

Figure 17.2 PDC Decision Tree with State-of-Nature Branch Probabilities



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 17.3 Applying the Expected Value Approach Using a Decision Tree for the PDC Condominium Project



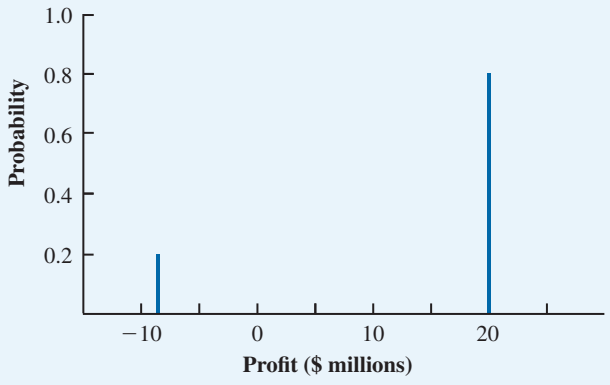
Risk Analysis

Risk analysis helps the decision maker recognize the difference between the expected value of a decision alternative and the payoff that may actually occur. A decision alternative and a state of nature combine to generate the payoff associated with a decision. The **risk profile** for a decision alternative shows the possible payoffs along with their associated probabilities.

Let us demonstrate risk analysis and the construction of a risk profile by returning to the PDC condominium construction project. Using the expected value approach, we identified the large condominium complex (d_3) as the best decision alternative. The expected value of \$14.2 million for d_3 is based on a 0.8 probability of obtaining a \$20 million profit and a 0.2 probability of obtaining a \$9 million loss. The 0.8 probability for the \$20 million payoff and the 0.2 probability for the −\$9 million payoff provide the risk profile for the large-complex decision alternative. This risk profile is shown graphically in Figure 17.4.

Sometimes a review of the risk profile associated with an optimal decision alternative may cause the decision maker to choose another decision alternative even though the

Figure 17.4 Risk Profile for the Large-Complex Decision Alternative for the PDC Condominium Project



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

expected value of the other decision alternative is not as good. For example, the risk profile for the medium-complex decision alternative (d_2) shows a 0.8 probability for a \$14 million payoff and a 0.2 probability for a \$5 million payoff. Because no probability of a loss is associated with decision alternative d_2, the medium-complex decision alternative would be judged less risky than the large-complex decision alternative. As a result, a decision maker might prefer the less risky medium-complex decision alternative even though it has an expected value of \$2 million less than the large-complex decision alternative.

Sensitivity Analysis

Sensitivity analysis can be used to determine how changes in the probabilities for the states of nature or changes in the payoffs affect the recommended decision alternative. In many cases, the probabilities for the states of nature and the payoffs are based on subjective assessments. Sensitivity analysis helps the decision maker understand which of these inputs are critical to the choice of the best decision alternative. If a small change in the value of one of the inputs causes a change in the recommended decision alternative, the solution to the decision analysis problem is sensitive to that particular input. Extra effort and care should be taken to make sure the input value is as accurate as possible. On the other hand, if a modest-to-large change in the value of one of the inputs does not cause a change in the recommended decision alternative, the solution to the decision analysis problem is not sensitive to that particular input. No extra time or effort would be needed to refine the estimated input value.

One approach to sensitivity analysis is to select different values for the probabilities of the states of nature and the payoffs and then resolve the decision analysis problem. If the recommended decision alternative changes, we know that the solution is sensitive to the changes made. For example, suppose that in the PDC problem the probability for a strong demand is revised to 0.2 and the probability for a weak demand is revised to 0.8. Would the recommended decision alternative change? Using $P(s_1) = 0.2$, $P(s_2) = 0.8$, and equation (17.2), the revised expected values for the three decision alternatives are as follows:

$$\begin{aligned} \text{EV}(d_1) &= 0.2(8) + 0.8(7) = 7.2 \\ \text{EV}(d_2) &= 0.2(14) + 0.8(5) = 6.8 \\ \text{EV}(d_3) &= 0.2(20) + 0.8(-9) = -3.2 \end{aligned}$$

With these probability assessments, the recommended decision alternative is to construct a small condominium complex (d_1), with an expected value of \$7.2 million. The probability of strong demand is only 0.2, so constructing the large condominium complex (d_3) is the least preferred alternative, with an expected value of −\$3.2 million (a loss).

Thus, when the probability of strong demand is large, PDC should build the large complex; when the probability of strong demand is small, PDC should build the small complex. Obviously, we could continue to modify the probabilities of the states of nature and learn even more about how changes in the probabilities affect the recommended decision alternative. Sensitivity analysis calculations can also be made for the values of the payoffs. We can easily change the payoff values and resolve the problem to see if the best decision changes.

Notes + Comments

1. The definition of expected value given in this chapter is consistent with that given in Chapter 5, but here we use the notation and terminology specific to decision analysis. In both cases, the expected value is defined as the weighted average of possible values.
2. The drawback to the sensitivity analysis approach described in this section is the numerous calculations required to evaluate the effect of several possible changes in the state-of-nature probabilities and/or payoff values. Many computer packages exist that can assist with the creation of decision trees and with the calculations required.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

17.4 Decision Analysis with Sample Information

Frequently, decision makers have the ability to collect additional information about the states of nature. It is worthwhile for the decision maker to consider the potential value of this additional information and how it can affect the decision analysis process. Most often, additional information is obtained through experiments designed to provide **sample information** about the states of nature. Raw material sampling, product testing, and market research studies are examples of experiments (or studies) that may enable management to revise or update the state-of-nature probabilities.

To analyze the potential benefit of additional information, we must first introduce a few additional terms related to decision analysis. The preliminary or **prior probabilities** assessments for the states of nature that are the best probability values available prior to obtaining additional information. These revised probabilities after obtaining additional information are called **posterior probabilities**.

Let us return to the PDC problem and assume that management is considering a six-month market research study designed to learn more about potential market acceptance of the PDC condominium project. Management anticipates that the market research study will provide one of the following two results:

1. *Favorable report:* A substantial number of the individuals contacted express interest in purchasing a PDC condominium.
2. *Unfavorable report:* Very few of the individuals contacted express interest in purchasing a PDC condominium.

The decision tree for the PDC problem with sample information shows the logical sequence for the decisions and the chance events in Figure 17.5. By introducing the possibility of conducting a market research study, the PDC problem becomes more complex. First, PDC's management must decide whether the market research should be conducted. If it is conducted, PDC's management must be prepared to make a decision about the size of the condominium project if the market research report is favorable and, possibly, a different decision about the size of the condominium project if the market research report is unfavorable. In Figure 17.5, the squares are decision nodes and the circles are chance nodes. At each decision node, the branch of the tree that is taken is based on the decision made. At each chance node, the branch of the tree that is taken is based on probability or chance. For example, decision node 1 shows that PDC must first make the decision of whether to conduct the market research study. If the market research study is undertaken, chance node 2 indicates that both the favorable report branch and the unfavorable report branch are not under PDC's control and will be determined by chance. Node 3 is a decision node, indicating that PDC must make the decision to construct the small, medium, or large complex if the market research report is favorable. Node 4 is a decision node showing that PDC must make the decision to construct the small, medium, or large complex if the market research report is unfavorable. Node 5 is a decision node indicating that PDC must make the decision to construct the small, medium, or large complex if the market research is not undertaken. Nodes 6 to 14 are chance nodes indicating that the strong demand or weak demand state-of-nature branches will be determined by chance.

Analysis of the decision tree and the choice of an optimal strategy require that we know the branch probabilities corresponding to all chance nodes. PDC has developed the following branch probabilities:

The branch probabilities for P(favorable report) and P(unfavorable report) are calculated using Bayes' theorem, first introduced in Chapter 5. We illustrate these calculations in Section 17.5.

If the market research study is undertaken:

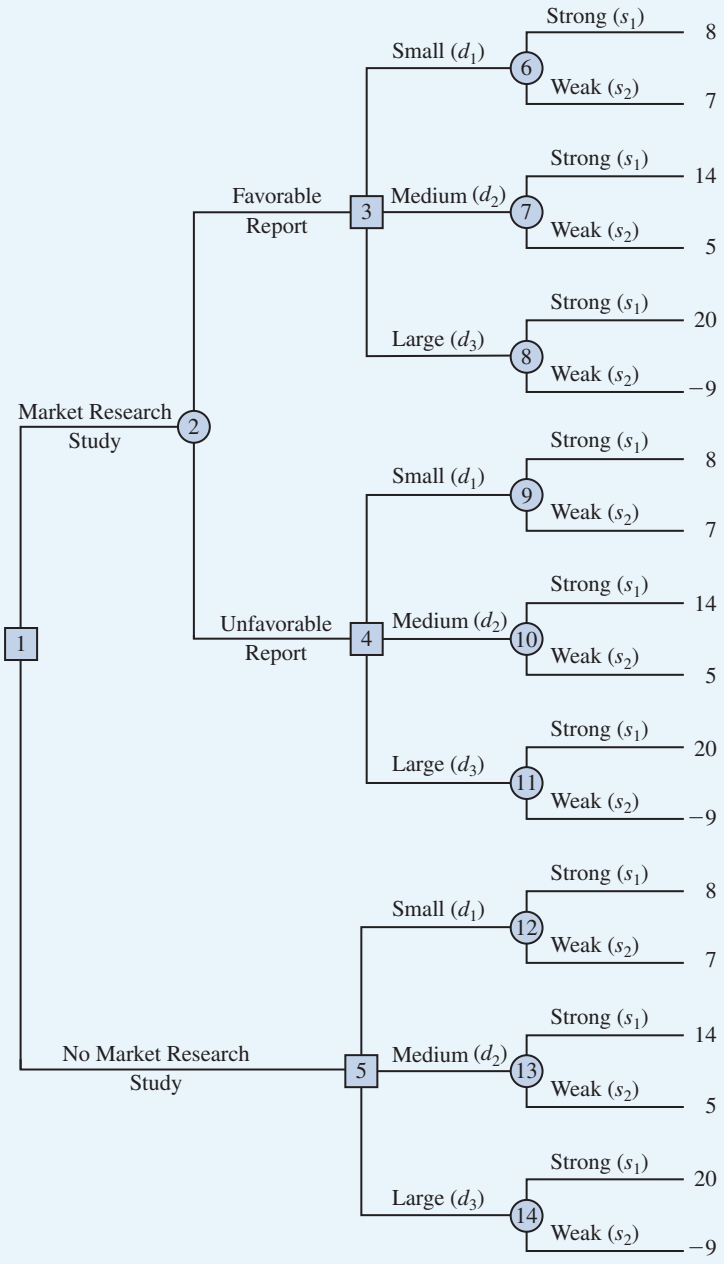
$$P(\text{favorable report}) = 0.77$$
$$P(\text{unfavorable report}) = 0.23$$

If the market research report is favorable, the posterior probabilities are as follows:

$$P(\text{strong demand given a favorable report}) = 0.94$$
$$P(\text{weak demand given a favorable report}) = 0.06$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 17.5 The PDC Decision Tree Including the Market Research Study



If the market research report is unfavorable, the posterior probabilities are as follows:

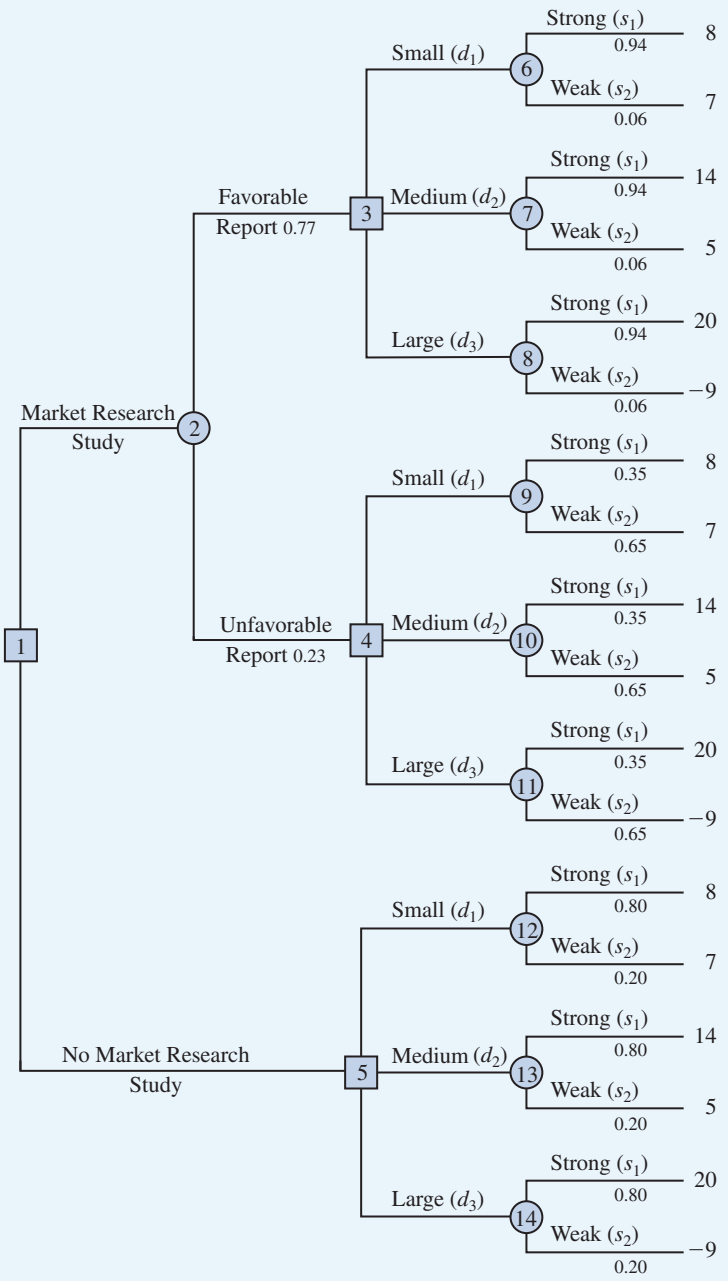
$$P(\text{strong demand given an unfavorable report}) = 0.35$$
$$P(\text{weak demand given an unfavorable report}) = 0.65$$

If the market research report is not undertaken, the prior probabilities are applicable:

$$P(\text{strong demand}) = 0.80$$
$$P(\text{weak demand}) = 0.20$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 17.6 The PDC Decision Tree with Branch Probabilities



The branch probabilities are shown on the decision tree in Figure 17.6.

A **decision strategy** is a sequence of decisions and chance outcomes in which the decisions chosen depend on the yet-to-be-determined outcomes of chance events. The approach used to determine the optimal decision strategy is based on a rollback of the expected values in the decision tree using the following steps:

1. At chance nodes, compute the expected value by multiplying the payoff at the end of each branch by the corresponding branch probabilities.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

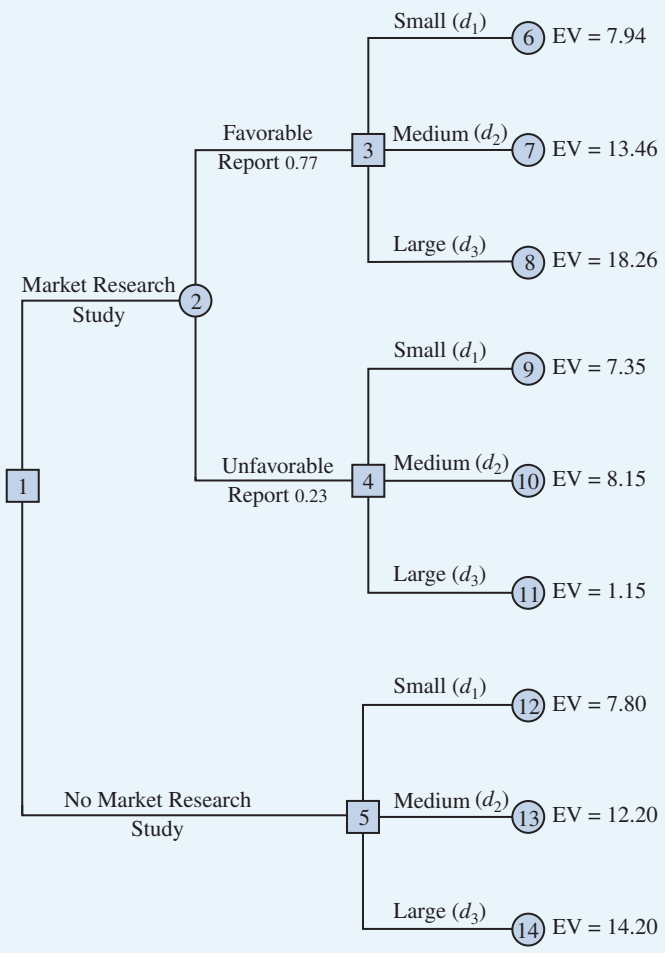
2. At decision nodes, select the decision branch that leads to the best expected value. This expected value becomes the expected value at the decision node.

Starting the rollback calculations by computing the expected values at chance nodes 6 to 14 provides the following results:

$$
\begin{aligned}
\text{EV(Node 6)} &= 0.94(8) + 0.06(7) &&= 7.94\\
\text{EV(Node 7)} &= 0.94(14) + 0.06(5) &&= 13.46\\
\text{EV(Node 8)} &= 0.94(20) + 0.06(-9) &&= 18.26\\
\text{EV(Node 9)} &= 0.35(8) + 0.65(7) &&= 7.35\\
\text{EV(Node 10)} &= 0.35(14) + 0.65(5) &&= 8.15\\
\text{EV(Node 11)} &= 0.35(20) + 0.65(-9) &&= 1.15\\
\text{EV(Node 12)} &= 0.80(8) + 0.20(7) &&= 7.80\\
\text{EV(Node 13)} &= 0.80(14) + 0.20(5) &&= 12.20\\
\text{EV(Node 14)} &= 0.80(20) + 0.20(-9) &&= 14.20
\end{aligned}
$$

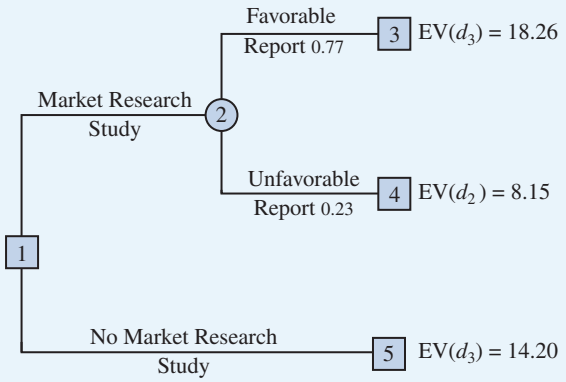
Figure 17.7 shows the reduced decision tree after computing expected values at these chance nodes.

Figure 17.7 PDC Decision Tree After Computing Expected Values at Chance Nodes 6 to 14



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 17.8 PDC Decision Tree After Choosing Best Decisions at Nodes 3, 4, and 5



Next, move to decision nodes 3, 4, and 5. For each of these nodes, we select the decision alternative branch that leads to the best expected value. For example, at node 3 we have the choice of the small complex branch with EV (Node 6) = 7.94, the medium complex branch with EV (Node 7) = 13.46, and the large complex branch with EV (Node 8) = 18.26. Thus, we select the large-complex decision alternative branch and the expected value at node 3 becomes EV (Node 3) = 18.26.

For node 4, we select the best expected value from nodes 9, 10, and 11. The best decision alternative is the medium complex branch that provides EV (Node 4) = 8.15. For node 5, we select the best expected value from nodes 12, 13, and 14. The best decision alternative is the large complex branch that provides EV (Node 5) = 14.20. Figure 17.8 shows the reduced decision tree after choosing the best decisions at nodes 3, 4, and 5 and rolling back the expected values to these nodes.

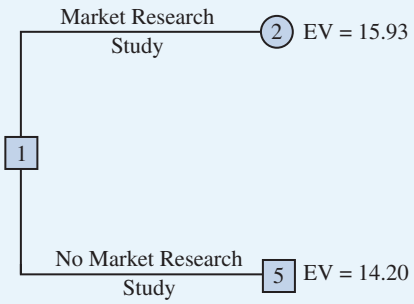
The expected value at chance node 2 can now be computed as follows:

$$\begin{aligned}\text{EV(Node 2)} &= 0.77\text{EV(Node 3)} + 0.23\text{EV(Node 4)} \\ &= 0.77(18.26) + 0.23(8.15) = 15.93\end{aligned}$$

This calculation reduces the decision tree to one involving only the two decision branches from node 1 (see Figure 17.9).

Finally, the decision can be made at decision node 1 by selecting the best expected values from nodes 2 and 5. This action leads to the decision alternative to conduct the market research study, which provides an overall expected value of 15.93.

Figure 17.9 PDC Decision Tree Reduced to Two Decision Branches



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The optimal decision for PDC is to conduct the market research study and then carry out the following decision strategy:

If the market research is favorable, construct the large condominium complex.
If the market research is unfavorable, construct the medium condominium complex.

The analysis of the PDC decision tree describes the methods that can be used to analyze more complex sequential decision problems. First, draw a decision tree consisting of decision and chance nodes and branches that describe the sequential nature of the problem. Determine the probabilities for all chance outcomes. Then, by working backward through the tree, compute expected values at all chance nodes and select the best decision branch at all decision nodes. The sequence of optimal decision branches determines the optimal decision strategy for the problem.

Expected Value of Sample Information

In the PDC problem, the market research study is the sample information used to determine the optimal decision strategy. The expected value associated with the market research study is 15.93. Previously, we showed that the best expected value if the market research study is *not* undertaken is 14.20. Thus, we can conclude that the difference, $15.93 - 14.20 = 1.73$, is the **expected value of sample information (EVSI)**. In other words, conducting the market research study adds \$1.73 million to the PDC expected value. The EVSI of \$1.73 million suggests PDC should be willing to pay up to \$1.73 million to conduct the market research study. In general, the expected value of sample information is as follows:

Expected Value of Sample Information (Evsi)

$$\text{EVSI} = |\text{EVwSI} - \text{EVwoSI}| \tag{17.3}$$

where

EVSI = expected value of sample information
EVwSI = expected value *with* sample information about the states of nature
EVwoSI = expected value *without* sample information about the states of nature

The absolute value in equation 17.3 allows this calculation to be correct for both minimization and maximization decision problems.

Expected Value of Perfect Information

A special case of gaining additional information related to a decision problem is when the sample information provides **perfect information** on the states of nature. In other words, consider a case in which the marketing study undertaken by PDC would determine exactly which state of nature will occur. Clearly, such a result is highly unlikely from a marketing study, but such an analysis provides a best-case analysis of the benefit provided by the marketing study. If the investment required for the additional information exceeds the expected value of perfect information, then we would not want to invest in procuring the additional information.

To illustrate the calculation of the expected value of perfect information, we return to the PDC decision. We assume for the moment that PDC could determine with certainty, prior to making a decision, which state of nature is going to occur. To make use of this perfect information, we will develop a decision strategy that PDC should follow once it knows which state of nature will occur.

To help determine the decision strategy for PDC, we reproduce PDC's payoff table as Table 17.6. If PDC knew for sure that state of nature s_1 would occur, the best decision alternative would be d_3, with a payoff of \$20 million. Similarly, if PDC knew for sure that state of nature s_2 would occur, the best decision alternative would be d_1, with a payoff of \$7 million. Thus, we can state PDC's optimal decision strategy when the perfect information becomes available as follows:

If s_1, select d_3 and receive a payoff of \$20 million.
If s_2, select d_1 and receive a payoff of \$7 million.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 17.6 Payoff Table for the PDC Condominium Project ($ millions)

Decision Alternative	State of Nature: Strong Demand, s_1	State of Nature: Weak Demand, s_2
Small complex, d_1	8	7
Medium complex, d_2	14	5
Large complex, d_3	20	−9

What is the expected value for this decision strategy? To compute the expected value with perfect information, we return to the original probabilities for the states of nature: $P(s_1) = 0.8$ and $P(s_2) = 0.2$. Thus, there is a 0.8 probability that the perfect information will indicate state of nature s_1, and the resulting decision alternative d_3 will provide a $20 million profit. Similarly, with a 0.2 probability for state of nature s_2, the optimal decision alternative d_1 will provide a $7 million profit. Thus, from equation (17.2) the expected value of the decision strategy that uses perfect information is $0.8(20) + 0.2(7) = 17.4$.

We refer to the expected value of $17.4 million as the *expected value with perfect information* (EVwPI).

Earlier in this section we showed that the recommended decision using the expected value approach is decision alternative d_3, with an expected value of $14.2 million. Because this decision recommendation and expected value computation were made without the benefit of perfect information, $14.2 million is referred to as the *expected value without perfect information* (EVwoPI).

The expected value with perfect information is $17.4 million, and the expected value without perfect information is $14.2; therefore, the expected value of the perfect information (EVPI) is $17.4 − $14.2 = $3.2 million. In other words, $3.2 million represents the additional expected value that can be obtained if perfect information were available about the states of nature. Therefore, it would be worth $3.2 million for PDC to learn the level of market acceptance before selecting a decision alternative. This represents the maximum that PDC should invest in any market research to provide additional information on the states of nature because no market research study can be expected to provide perfect information.

In general, the **expected value of perfect information (EVPI)** is computed as shown in equation (17.4).

Expected Value of Perfect Information (EVPI)

$$\text{EVPI} = |\text{EVwPI} - \text{EVwoPI}| \tag{17.4}$$

where

EVPI = expected value of perfect information

EVwPI = expected value *with* perfect information about the states of nature

EVwoPI = expected value *without* perfect information about the states of nature

17.5 Computing Branch Probabilities with Bayes' Theorem

We first introduced Bayes' theorem in Chapter 5 as a means of calculating posterior probabilities as updates of prior probabilities once additional information is obtained.

In Section 17.4, the branch probabilities for the PDC decision tree chance nodes were provided in the problem description. No computations were required to determine these probabilities. In this section, we show how **Bayes' theorem** can be used to compute branch probabilities for decision trees. The branch probabilities are the posterior probabilities for demand that have been updated based on the sample information of whether the market research report is favorable or unfavorable.

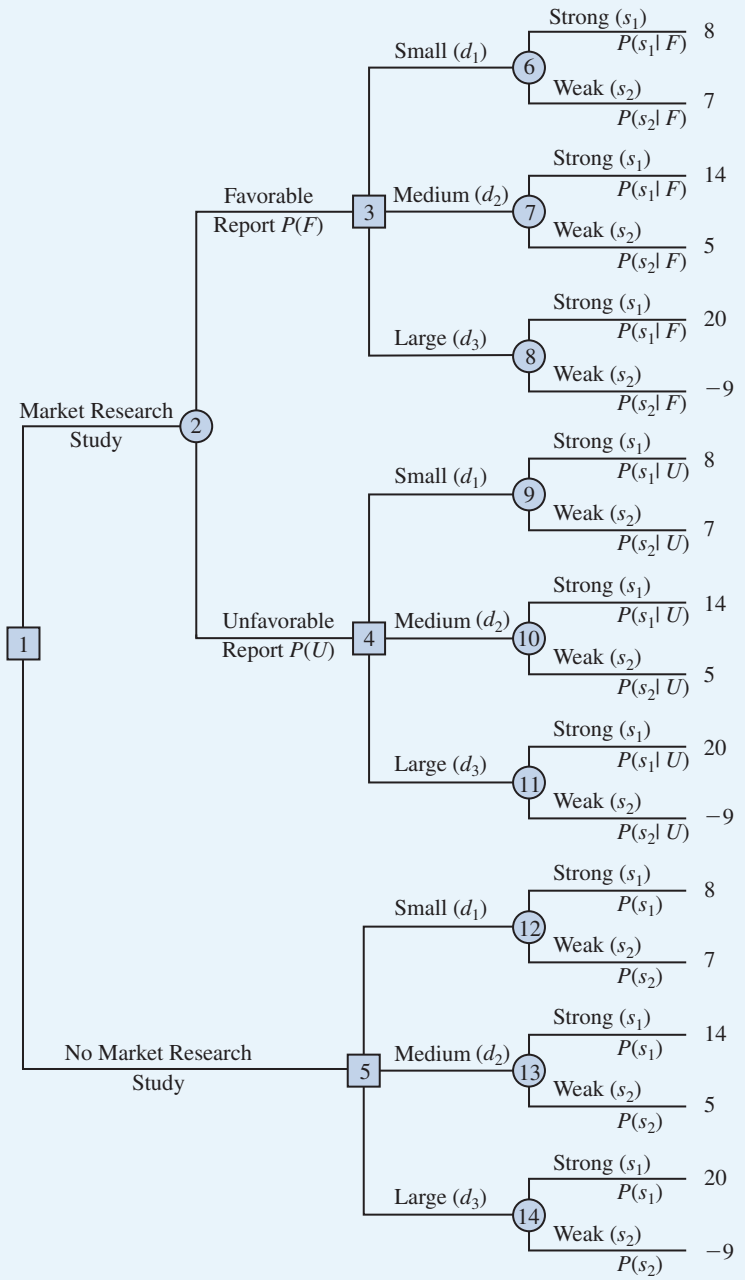
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The PDC decision tree is shown again in Figure 17.10. Let

F = favorable market research report
U = unfavorable market research report
s_1 = strong demand (state of nature 1)
s_2 = weak demand (state of nature 2)

At chance node 2, we need to know the branch probabilities $P(F)$ and $P(U)$. At chance nodes 6, 7, and 8, we need to know the branch probabilities $P(s_1 | F)$, which is read as "the probability of state of nature 1 given a favorable market research report," and $P(s_2 | F)$,

Figure 17.10 The PDC Decision Tree



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Conditional probability is introduced in Chapter 5.

which is the probability of state of nature 2 given a favorable market research report. The notation | in $P(s_1|F)$ and $P(s_2|F)$ is read as "given" and indicates a **conditional probability**, because we are interested in the probability of a particular state of nature "conditioned" on the fact that we receive a favorable market report. $P(s_1|F)$ and $P(s_2|F)$ are referred to as *posterior probabilities* because they are conditional probabilities based on the outcome of the sample information. At chance nodes 9, 10, and 11, we need to know the branch probabilities $P(s_1|U)$ and $P(s_2|U)$; note that these are also posterior probabilities, denoting the probabilities of the two states of nature *given* that the market research report is unfavorable. Finally, at chance nodes 12, 13, and 14, we need the probabilities for the states of nature, $P(s_1)$ and $P(s_2)$, if the market research study is not undertaken.

In performing the probability computations, we need to know PDC's assessment of the probabilities for the two states of nature, $P(s_1)$ and $P(s_2)$, which are the prior probabilities as discussed earlier. In addition, we must know the conditional probability of the market research outcomes (the sample information) *given* each state of nature. For example, we need to know the conditional probability of a favorable market research report given that the state of nature is strong demand for the PDC project. To carry out the probability calculations, we will need conditional probabilities for all sample outcomes given all states of nature, that is, $P(F|s_1)$, $P(F|s_2)$, $P(U|s_1)$, and $P(U|s_2)$. These conditional probabilities are assessments of the accuracy of the market research; they are often estimated using historical performance of previous market research reports. For example, $P(F|s_1)$ may be estimated via the historical frequency of strong demand being associated with a market research report that was favorable. In the PDC problem, we assume that the following assessments are available for these conditional probabilities:

	Market Research	
State of Nature	**Favorable, *F***	**Unfavorable, *U***
Strong demand, s_1	$P(F\|s_1) = 0.90$	$P(U\|s_1) = 0.10$
Weak demand, s_2	$P(F\|s_2) = 0.25$	$P(U\|s_2) = 0.75$

Note that the preceding probability assessments provide a reasonable degree of confidence in the market research study. If the true state of nature is s_1, the probability of a favorable market research report is 0.90, and the probability of an unfavorable market research report is 0.10. If the true state of nature is s_2, the probability of a favorable market research report is 0.25, and the probability of an unfavorable market research report is 0.75. One reason for a 0.25 probability of a potentially misleading favorable market research report for state of nature s_2 is that when some potential buyers first hear about the new condominium project, their enthusiasm may lead them to overstate their real interest in it. A potential buyer's initial favorable response can change quickly to a "no-thank-you" when later faced with the reality of signing a purchase contract and making a down payment.

Equation (17.5) is a restatement of Bayes' theorem introduced in Chapter 5.

Equation (17.5) is known as Bayes' theorem, and it is used to compute posterior probabilities.

Bayes' Theorem

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_n)} \qquad \textbf{(17.5)}$$

To perform the Bayes' theorem calculations for $P(s_1|U)$ using equation (17.5), we replace B with U (unfavorable report) and A_i with s_1 in equation (17.5) so that we have

$$P(s_1|U) = \frac{P(U|s_1)P(s_1)}{P(U|s_1)P(s_1) + P(U|s_2)P(s_2)}$$

$$= \frac{0.10 \times 0.80}{(0.10 \times 0.80) + (0.20 \times 0.75)} = 0.35$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

which indicates that the probability of strong demand given an unfavorable market research report is 0.35. We can also calculate the probability of weak demand given an unfavorable market research report as shown below:

$$P(s_2|U) = \frac{P(U|s_2)P(s_2)}{P(U|s_1)P(s_1) + P(U|s_2)P(s_2)} = \frac{0.75 \times 0.20}{(0.10 \times 0.80) + (0.75 \times 0.20)} = 0.65$$

Similarly, we can calculate the posterior probabilities for strong and weak demand given a favorable market research report using equation (17.5):

$$P(s_1|F) = \frac{P(F|s_1)P(s_1)}{P(F|s_1)P(s_1) + P(F|s_2)P(s_2)} = \frac{0.90 \times 0.80}{(0.90 \times 0.80) + (0.25 \times 0.20)} = 0.94$$

and

$$P(s_2|F) = \frac{P(F|s_2)P(s_2)}{P(F|s_1)P(s_1) + P(F|s_2)P(s_2)} = \frac{0.25 \times 0.20}{(0.90 \times 0.80) + (0.25 \times 0.20)} = 0.06$$

This indicates that a favorable research report leads to a posterior probability of 0.94 that the demand will be strong and a posterior probability of only 0.06 that the demand will be weak.

The discussion in this section shows an underlying relationship between the probabilities on the various branches in a decision tree. It would be inappropriate to assume different prior probabilities, $P(s_1)$ and $P(s_2)$, without determining how these changes would alter $P(F)$ and $P(U)$, as well as the posterior probabilities $P(s_1|F)$, $P(s_2|F)$, $P(s_1|U)$, and $P(s_2|U)$.

17.6 Utility Theory

The decision analysis situations presented so far in this chapter expressed outcomes (payoffs) in terms of monetary values. With probability information available about the outcomes of the chance events, we defined the optimal decision alternative as the one that provides the best expected value. However, in some situations the decision alternative with the best expected value may not be the preferred alternative. A decision maker may also wish to consider intangible factors such as risk, image, or other nonmonetary criteria in order to evaluate the decision alternatives. When monetary value does not necessarily lead to the most preferred decision, expressing the value (or worth) of a consequence in terms of its utility will permit the use of expected utility to identify the most desirable decision alternative. The discussion of utility and its application in decision analysis is presented in this section.

Utility is a measure of the total worth or relative desirability of a particular outcome; it reflects the decision maker's attitude toward a collection of factors such as profit, loss, and risk. Researchers have found that as long as the monetary value of payoffs stays within a range that the decision maker considers reasonable, selecting the decision alternative with the best expected value usually leads to selection of the most preferred decision. However, when the payoffs are extreme, decision makers are often unsatisfied or uneasy with the decision that simply provides the best expected value.

As an example of a situation in which utility can help in selecting the best decision alternative, let us consider the problem faced by Swofford, Inc., a relatively small real estate investment firm located in Atlanta, Georgia. Swofford currently has two investment opportunities that require approximately the same cash outlay. The cash requirements necessary prohibit Swofford from making more than one investment at this time. Consequently, three possible decision alternatives may be considered.

The three decision alternatives, denoted by d_1, d_2, and d_3, are as follows:

d_1 = make investment A
d_2 = make investment B
d_3 = do not invest

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 17.7 Payoff Table for Swofford, Inc.

	State of Nature		
Decision Alternative	**Prices Up, s_1**	**Prices Stable, s_2**	**Prices Down, s_3**
Investment A, d_1	$30,000	$20,000	−$50,000
Investment B, d_2	$50,000	−$20,000	−$30,000
Do not invest, d_3	$0	$0	$0

The monetary payoffs associated with the investment opportunities depend on the investment decision and on the direction of the real estate market during the next six months (the chance event). Real estate prices will go up, remain stable, or go down. Thus, the states of nature, denoted by s_1, s_2, and s_3, are as follows:

$$s_1 = \text{real estate prices go up}$$
$$s_2 = \text{real estate prices remain stable}$$
$$s_3 = \text{real estate prices go down}$$

Using the best information available, Swofford has estimated the profits, or payoffs, associated with each decision alternative and state-of-nature combination. The resulting payoff table is shown in Table 17.7.

The best estimate of the probability that real estate prices will go up is 0.3; the best estimate of the probability that prices will remain stable is 0.5; and the best estimate of the probability that prices will go down is 0.2. Thus, the expected values for the three decision alternatives are as follows:

$$\begin{aligned} EV(d_1) &= 0.3(30,000) + 0.5(20,000) + 0.2(-50,000) = 9,000 \\ EV(d_2) &= 0.3(50,000) + 0.5(-20,000) + 0.2(-30,000) = -11,000 \\ EV(d_3) &= 0.3(0) + 0.5(0) + 0.2(0) = 0 \end{aligned}$$

Using the expected value approach, the optimal decision is to select investment A with an expected value of $9,000. Is it really the best decision alternative? Let us consider some other relevant factors that relate to Swofford's capability for absorbing the loss of $50,000 if investment A is made and prices actually go down.

Actually, Swofford's current financial position is weak. This condition is partly reflected in Swofford's ability to make only one investment. More important, however, the firm's president believes that, if the next investment results in a substantial loss, Swofford's future will be in jeopardy. Although the expected value approach leads to a recommendation for d_1, do you think the firm's president would prefer this decision? We suspect that the president would select d_2 or d_3 to avoid the possibility of incurring a $50,000 loss. In fact, a reasonable conclusion is that, if a loss of even $30,000 could drive Swofford out of business, the president would select d_3, believing that both investments A and B are too risky for Swofford's current financial position.

The way we resolve Swofford's dilemma is first to determine Swofford's utility for the various outcomes. Recall that the utility of any outcome is the total worth of that outcome, taking into account all risks and consequences involved. If the utilities for the various consequences are assessed correctly, the decision alternative with the highest expected utility is the most preferred, or best, alternative. We next show how to determine the utility of the outcomes so that the alternative with the highest expected utility can be identified.

Utility and Decision Analysis

The procedure we use to establish a utility for each of the payoffs in Swofford's situation requires that we first assign a utility to the best and worst possible payoffs. Any values will

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

work as long as the utility assigned to the best payoff is greater than the utility assigned to the worst payoff. In this case, \$50,000 is the best payoff and −\$50,000 is the worst. Suppose, then, that we arbitrarily make assignments to these two payoffs as follows:

Utility values of 0 and 1 could be used here. We select 0 and 10 to avoid any possible confusion between the utility value for a payoff and the probability p.

$$\text{Utility of } -\$50{,}000 = U(-50{,}000) = 0$$
$$\text{Utility of } \$50{,}000 = U(50{,}000) = 10$$

Let us now determine the utility associated with every other payoff.

Consider the process of establishing the utility of a payoff of \$30,000. First we ask Swofford's president to state a preference between a guaranteed \$30,000 payoff and an opportunity to engage in the following lottery, or bet, for some probability of p that we select:

p *is often referred to as the indifference probability.*

Lottery: Swofford obtains a payoff of \$50,000 with probability p and a payoff of −\$50,000 with probability $(1 - p)$.

Obviously, if p is very close to 1, Swofford's president would prefer the lottery to the guaranteed payoff of \$30,000 because the firm would virtually ensure itself a payoff of \$50,000. If p is very close to 0, Swofford's president would clearly prefer the guarantee of \$30,000. In any event, as p increases continuously from 0 to 1, the preference for the guaranteed payoff of \$30,000 decreases and at some point is equal to the preference for the lottery. At this value of p, Swofford's president would have equal preference for the guaranteed payoff of \$30,000 and the lottery; at greater values of p, Swofford's president would prefer the lottery to the guaranteed \$30,000 payoff. For example, let us assume that when $p = 0.95$, Swofford's president is indifferent between the guaranteed payoff of \$30,000 and the lottery. For this value of p, we can compute the utility of a \$30,000 payoff as follows:

$$\begin{aligned} U(30{,}000) &= pU(50{,}000) + (1-p)U(-50{,}000) \\ &= 0.95(10) + (0.05)(0) \\ &= 9.5 \end{aligned}$$

Obviously, if we had started with a different assignment of utilities for a payoff of \$50,000 and −\$50,000, the result would have been a different utility for \$30,000. For example, if we had started with an assignment of 100 for \$50,000 and 10 for −\$50,000, the utility of a \$30,000 payoff would be

$$\begin{aligned} U(30{,}000) &= 0.95(100) + 0.05(10) \\ &= 95.0 + 0.5 \\ &= 95.5 \end{aligned}$$

Hence, we must conclude that the utility assigned to each payoff is not unique but merely depends on the initial choice of utilities for the best and worst payoffs.

Before computing the utility for the other payoffs, let us consider the implication of Swofford's president assigning a utility of 9.5 to a payoff of \$30,000. Clearly, when $p = 0.95$, the expected value of the lottery is

$$\begin{aligned} \text{EV(lottery)} &= 0.95(\$50{,}000) + 0.05(-\$50{,}000) \\ &= \$47{,}500 - \$2{,}500 \\ &= \$45{,}000 \end{aligned}$$

The difference between the expected value of the lottery and the guaranteed payoff can be viewed as the risk premium the decision maker is willing to pay.

Although the expected value of the lottery when $p = 0.95$ is \$45,000, Swofford's president is indifferent between the lottery (and its associated risk) and a guaranteed payoff of \$30,000. Thus, Swofford's president is taking a conservative, or risk-avoiding, viewpoint. A decision maker who would choose a guaranteed payoff over a lottery with a superior expected payoff is a **risk avoider** (or is said to be risk-averse). The president would rather have \$30,000 for certain than risk anything greater than a 5% chance of incurring a loss of \$50,000. In other words, the difference between the EV of \$45,000 and the guaranteed

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

payoff of \$30,000 is the risk premium that Swofford's president would be willing to pay to avoid the 5% chance of losing \$50,000.

To compute the utility associated with a payoff of −\$20,000, we must ask Swofford's president to state a preference between a guaranteed −\$20,000 payoff and an opportunity to engage again in the following lottery:

Lottery: Swofford obtains a payoff of \$50,000 with probability p and a payoff of −\$50,000 with probability $(1 - p)$.

Note that this lottery is exactly the same as the one we used to establish the utility of a payoff of \$30,000 (in fact, we can use this lottery to establish the utility for any value in the Swofford payoff table). We need to determine the value of p that would make the president indifferent between a guaranteed payoff of −\$20,000 and the lottery. For example, we might begin by asking the president to choose between a certain loss of \$20,000 and the lottery with a payoff of \$50,000 with probability $p = 0.90$ and a payoff of −\$50,000 with probability $(1 - p) = 0.10$. What answer do you think we would get? Surely, with this high probability of obtaining a payoff of \$50,000, the president would elect the lottery. Next, we might ask whether $p = 0.85$ would result in indifference between the loss of \$20,000 for certain and the lottery. Again the president might prefer the lottery. Suppose that we continue until we get to $p = 0.55$, at which point the president is indifferent between the payoff of −\$20,000 and the lottery. In other words, for any value of p less than 0.55, the president would take a loss of \$20,000 for certain rather than risk the potential loss of \$50,000 with the lottery; and for any value of p above 0.55, the president would choose the lottery. Thus, the utility assigned to a payoff of −\$20,000 is

$$\begin{aligned} U(-\$20{,}000) &= pU(50{,}000) + (1 - p)U(-\$50{,}000) \\ &= 0.55(10) + 0.45(0) \\ &= 5.5 \end{aligned}$$

Again let us assess the implication of this assignment by comparing it to the expected value approach. When $p = 0.55$, the expected value of the lottery is

$$\begin{aligned} \text{EV(lottery)} &= 0.55(\$50{,}000) + 0.45(-\$50{,}000) \\ &= \$27{,}500 - \$22{,}500 \\ &= \$5{,}000 \end{aligned}$$

Thus, Swofford's president would just as soon absorb a certain loss of \$20,000 as take the lottery and its associated risk, even though the expected value of the lottery is \$5,000. Once again this preference demonstrates the conservative, or risk-avoiding, point of view of Swofford's president.

In these two examples, we computed the utility for the payoffs of \$30,000 and −\$20,000. We can determine the utility for any payoff M in a similar fashion. First, we must find the probability p for which the decision maker is indifferent between a guaranteed payoff of M and a lottery with a payoff of \$50,000 with probability p and −\$50,000 with probability $(1 - p)$. The utility of M is then computed as follows:

$$\begin{aligned} U(M) &= pU(\$50{,}000) + (1 - p)U(-\$50{,}000) \\ &= p(10) + (1 - p)0 \\ &= 10p \end{aligned}$$

Using this procedure we developed a utility for each of the remaining payoffs in Swofford problem. The results are presented in Table 17.8.

Now that we have determined the utility of each of the possible monetary values, we can write the original payoff table in terms of utility. Table 17.9 shows the utility for the various outcomes in the Swofford problem. The notation we use for the entries in the utility table is U_{ij}, which denotes the utility associated with decision alternative d_i and state of nature s_j. Using this notation, we see that $U_{23} = 4.0$.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 17.8 Utility of Monetary Payoffs for Swofford, Inc.

Monetary Value ($)	Indifference Value of p	Utility
50,000	Does not apply	10.0
30,000	0.95	9.5
20,000	0.90	9.0
0	0.75	7.5
−20,000	0.55	5.5
−30,000	0.40	4.0
−50,000	Does not apply	0

Table 17.9 Utility Table for Swofford, Inc.

	State of Nature		
Decision Alternative	Prices Up, s_1	Prices Stable, s_2	Prices Down, s_3
Investment A, d_1	9.5	9.0	0
Investment B, d_2	10.0	5.5	4.0
Do not invest, d_3	7.5	7.5	7.5

We can now compute the **expected utility (EU)** of the utilities in Table 17.9 in a similar fashion as we computed expected value in Section 17.3. In other words, to identify an optimal decision alternative for Swofford, Inc., the expected utility approach requires the analyst to compute the expected utility for each decision alternative and then select the alternative yielding the highest expected utility. With N possible states of nature, the expected utility of a decision alternative d_i is given as follows:

Expected Utility (EU)

$$EU(d_i) = \sum_{j=1}^{N} P(s_j)U_{ij} \tag{17.6}$$

The expected utility for each of the decision alternatives in the Swofford problem is as follows:

$$\begin{aligned} \text{EU}(d_1) &= 0.3(9.5) + 0.5(9.0) + 0.2(0) \quad = 7.35 \\ \text{EU}(d_2) &= 0.3(10) \ + 0.5(5.5) + 0.2(4.0) = 6.55 \\ \text{EU}(d_3) &= 0.3(7.5) + 0.5(7.5) + 0.2(7.5) = 7.50 \end{aligned}$$

Note that the optimal decision using the expected utility approach is d_3; do not invest. The ranking of alternatives according to the president's utility assignments and the associated monetary values are as follows:

Ranking of Decision Alternatives	Expected Utility	Expected Value ($)
Do not invest	7.50	0
Investment A	7.35	9,000
Investment B	6.55	−1,000

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Note that, although investment A had the highest expected value of $9,000, the analysis indicates that Swofford should decline this investment. The rationale behind not selecting investment A is that the 0.20 probability of a $50,000 loss was considered by Swofford's president to involve a serious risk. The seriousness of this risk and its associated impact on the company were not adequately reflected by the expected value of investment A. We assessed the utility for each payoff to assess this risk adequately.

The following steps state in general terms the procedure used to solve the Swofford, Inc. investment problem:

Step 1. Develop a payoff table using monetary values

Step 2. Identify the best and worst payoff values in the table and assign each a utility, with $U(\text{best payoff}) > U(\text{worst payoff})$

Step 3. For every other monetary value M in the original payoff table, do the following to determine its utility:

a. Define the lottery such that there is a probability p of the best payoff and a probability $(1 - p)$ of the worst payoff

b. Determine the value of p such that the decision maker is indifferent between a guaranteed payoff of M and the lottery defined in Step 3(a)

c. Calculate the utility of M as follows:

$$U(M) = pU\big(\text{best payoff}\big) + (1 - p)U\big(\text{worst payoff}\big)$$

Step 4. Convert each monetary value in the payoff table to a utility

Step 5. Apply the expected utility approach to the utility table developed in Step 4 and select the decision alternative with the highest expected utility

The procedure we described for determining the utility of monetary consequences can also be used to develop a utility measure for nonmonetary consequences. Assign the best consequence a utility of 10 and the worst a utility of 0. Then create a lottery with a probability of p for the best consequence and $(1 - p)$ for the worst consequence. For each of the other consequences, find the value of p that makes the decision maker indifferent between the lottery and the consequence. Then calculate the utility of the consequence in question as follows:

$$U(\text{consequence}) = pU(\text{best consequence}) + (1 - p)U(\text{worst consequence})$$

Utility Functions

Next, we describe how different decision makers may approach risk in terms of their assessment of utility. The financial position of Swofford, Inc. was such that the firm's president evaluated investment opportunities from a conservative, or risk-avoiding, point of view. However, if the firm had a surplus of cash and a stable future, Swofford's president might have been looking for investment alternatives that, although perhaps risky, contained a potential for substantial profit. That type of behavior would demonstrate that the president is a risk taker with respect to this decision.

A **risk taker** is a decision maker who would choose a lottery over a guaranteed payoff when the expected value of the lottery is inferior to the guaranteed payoff. In this section, we analyze the decision problem faced by Swofford from the point of view of a decision maker who would be classified as a risk taker. We then compare the conservative point of view of Swofford's president (a risk avoider) with the behavior of a decision maker who is a risk taker.

For the decision problem facing Swofford, Inc., using the general procedure for developing utilities as discussed previously, a risk taker might express the utility for the various payoffs shown in Table 17.10. As before, $U(50{,}000) = 10$ and $U(-50{,}000) = 0$. Note the difference in behavior reflected in Tables 17.10 and 17.8. In other words, in determining the value of p at which the decision maker is indifferent between a guaranteed payoff of M and a lottery in which $50,000 is obtained with probability p and $-$$50,000 with probability $(1 - p)$, the risk taker is willing to accept a greater risk of incurring a loss of $50,000 in order to gain the opportunity to realize a profit of $50,000.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table 17.10 Revised Utilities for Swofford, Inc., Assuming a Risk Taker

Monetary Value ($)	Indifference Value of p	Utility
50,000	Does not apply	10.0
30,000	0.50	5.0
20,000	0.40	4.0
0	0.25	2.5
−20,000	0.15	1.5
−30,000	0.10	1.0
−50,000	Does not apply	0

To help develop the utility table for the risk taker, we have reproduced the Swofford, Inc. payoff table in Table 17.11. Using these payoffs and the risk taker's utilities given in Table 17.10, we can write the risk taker's utility table as shown in Table 17.12. Using the state-of-nature probabilities $P(s_1) = 0.3$, $P(s_2) = 0.5$, and $P(s_3) = 0.2$, the expected utility for each decision alternative is as follows:

$$\begin{aligned}
\text{EU}(d_1) &= 0.3(5.0) + 0.5(4.0) + 0.2(0) &&= 3.50\\
\text{EU}(d_2) &= 0.3(10) + 0.5(1.5) + 0.2(1.0) &&= 3.95\\
\text{EU}(d_3) &= 0.3(2.5) + 0.5(2.5) + 0.2(2.5) &&= 2.50
\end{aligned}$$

What is the recommended decision? Perhaps somewhat to your surprise, the analysis recommends investment B, with the highest expected utility of 3.95. Recall that this investment has a −$1,000 expected value. Why is it now the recommended decision? Remember that the decision maker in this revised problem is a risk taker. Thus, although the expected value of investment B is negative, utility analysis has shown that this decision maker is enough of a risk taker to prefer investment B and its potential for the $50,000 profit.

Table 17.11 Payoff Table for Swofford, Inc.

	State of Nature		
Decision Alternative	Prices Up, s_1	Prices Stable, s_2	Prices Down, s_3
Investment A, d_1	$30,000	$20,000	−$50,000
Investment B, d_2	$50,000	−$20,000	−$30,000
Do not invest, d_3	$0	$0	$0

Table 17.12 Utility Table of a Risk Taker for Swofford, Inc.

	State of Nature		
Decision Alternative	Prices Up, s_1	Prices Stable, s_2	Prices Down, s_3
Investment A, d_1	5.0	4.0	0
Investment B, d_2	10.0	1.5	1.0
Do not invest, d_3	2.5	2.5	2.5

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Ranking by the expected utilities generates the following order of preference of the decision alternatives for the risk taker and the associated expected values:

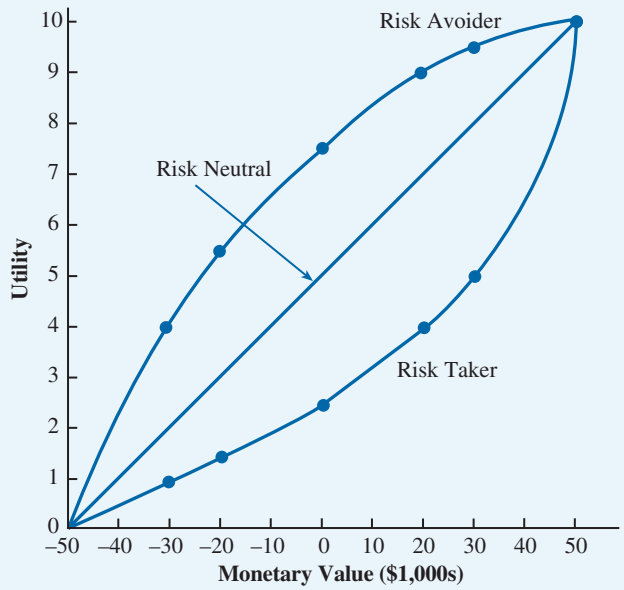
Ranking of Decision Alternatives	Expected Utility	Expected Value ($)
Investment B	3.95	−1,000
Investment A	3.50	9,000
Do not invest	2.50	0

Comparing the utility analysis for a risk taker with the more conservative preferences of the president of Swofford, Inc., who is a risk avoider, we see that, even with the same decision problem, different attitudes toward risk can lead to different recommended decisions. The utilities established by Swofford's president indicated that the firm should not invest at this time, whereas the utilities established by the risk taker showed a preference for investment B. Note that both of these decisions differ from the best expected value decision, which was investment A.

We can obtain another perspective of the difference between behaviors of a risk avoider and a risk taker by developing a graph that depicts the relationship between monetary value and utility. We use the horizontal axis of the graph to represent monetary values and the vertical axis to represent the utility associated with each monetary value. Now, consider the data in Table 17.8, with a utility corresponding to each monetary value for the original Swofford, Inc. problem. These values can be plotted on a graph to produce the top curve in Figure 17.11. The resulting curve is the **utility function for money** for Swofford's president. Recall that these points reflected the conservative, or risk-avoiding, nature of Swofford's president. Hence, we refer to the top curve in Figure 17.11 as a utility function for a risk avoider. Using the data in Table 17.10 developed for a risk taker, we can plot these points to produce the bottom curve in Figure 17.11. The resulting curve depicts the utility function for a risk taker.

By looking at the utility functions in Figure 17.11, we can begin to generalize about the utility functions for risk avoiders and risk takers. Although the exact shape of the utility

Figure 17.11 Utility Function for Money for Risk-Avoider, Risk-Taker, and Risk-Neutral Decision Makers



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

function will vary from one decision maker to another, we can see the general shape of these two types of utility functions. The utility function for a risk avoider shows a diminishing marginal return for money. For example, the increase in utility going from a monetary value of −$30,000 to $0 is 7.5 − 4.0 = 3.5, whereas the increase in utility in going from $0 to $30,000 is only 9.5 − 7.5 = 2.0.

However, the utility function for a risk taker shows an increasing marginal return for money. For example, in Figure 17.11, the increase in utility for the risk taker in going from −$30,000 to $0 is 2.5 − 1.0 = 1.5, whereas the increase in utility in going from $0 to $30,000 for the risk taker is 5.0 − 2.5 = 2.5. Note also that in either case the utility function is always increasing; that is, more money leads to more utility. All utility functions possess this property.

We concluded that the utility function for a risk avoider shows a diminishing marginal return for money and that the utility function for a risk taker shows an increasing marginal return. When the marginal return for money is neither decreasing nor increasing but remains constant, the corresponding utility function describes the behavior of a decision maker who is neutral to risk. The following characteristics are associated with a **risk-neutral** decision maker:

1. The utility function can be drawn as a straight line connecting the "best" and the "worst" points.
2. The expected utility approach and the expected value approach applied to monetary payoffs result in the same action.

The straight, diagonal line in Figure 17.11 depicts the utility function of a risk-neutral decision maker using the Swofford, Inc. problem data.

Generally, when the payoffs for a particular decision-making problem fall into a reasonable range—the best is not too good and the worst is not too bad—decision makers tend to express preferences in agreement with the expected value approach. Thus, we suggest asking the decision maker to consider the best and worst possible payoffs for a problem and assess their reasonableness. If the decision maker believes that they are in the reasonable range, the decision alternative with the best expected value can be used. However, if the payoffs appear unreasonably large or unreasonably small (e.g., a huge loss) and if the decision maker believes that monetary values do not adequately reflect their true preferences for the payoffs, a utility analysis of the problem should be considered.

Exponential Utility Function

Having a decision maker provide enough indifference values to create a utility function can be time consuming. An alternative is to assume that the decision maker's utility is defined by an exponential function. Figure 17.12 shows examples of different exponential utility functions. Note that all the exponential utility functions indicate that the decision maker is risk averse. The form of the exponential utility function is as follows:

In equation (17.7), the number e ≈ 2.718282… is a mathematical constant corresponding to the base of the natural logarithm. In Excel, e^x can be evaluated for any power x using the function EXP(x).

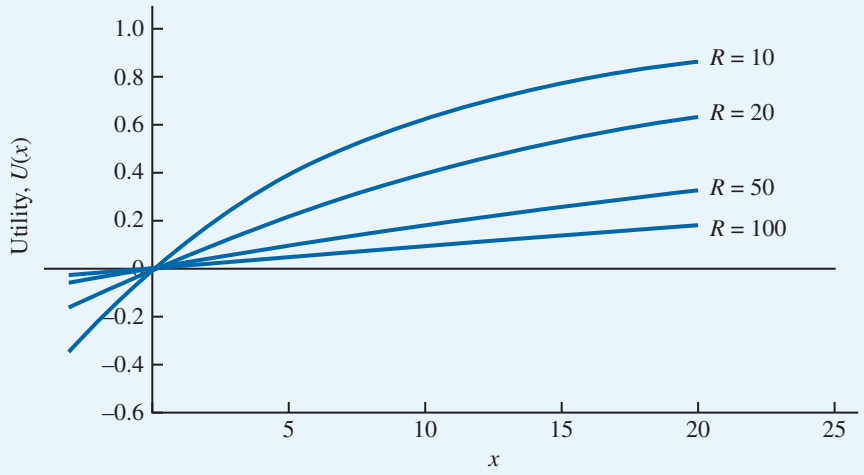
Exponential Utility Function

$$U(x) = 1 - e^{-x/R} \tag{17.7}$$

The R parameter in equation (17.7) represents the decision maker's risk tolerance; it controls the shape of the exponential utility function. Larger R values create flatter exponential functions, indicating that the decision maker is less risk averse (closer to risk neutral). Smaller R values indicate that the decision maker has less risk tolerance (is more risk averse). A common method to determine an approximate risk tolerance is to ask the decision maker to consider a scenario in which they could win $\$R$ with probability 0.5 and lose $\$R/2$ with probability 0.5. The R value to use in equation (17.7) is the largest $\$R$ for which the decision maker would accept this gamble. For instance, if the decision maker is comfortable accepting

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure 17.12 Exponential Utility Functions with Different Risk Tolerance (R) Values



a gamble with a 50% chance of winning \$2,000 and a 50% chance of losing \$1,000, but not with a gamble with a 50% chance of winning \$3,000 and a 50% chance of losing \$1,500, then we would use $R = \$2{,}000$ in equation (17.7). Determining the maximum gamble that a decision maker is willing to take and then using this value in the exponential utility function can be much less time consuming than generating a complete table of indifference probabilities. One should remember that using an exponential utility function assumes that the decision maker is risk averse; however, this is often true in practice for business decisions.

Notes + Comments

1. In the Swofford problem, we have been using a utility of 10 for the best payoff and 0 for the worst. We could have chosen any values as long as the utility associated with the best payoff exceeds the utility associated with the worst payoff. Alternatively, a utility of 1 can be associated with the best payoff and a utility of 0 can be associated with the worst payoff. Had we made this choice, the utility for any monetary value M would have been the value of p at which the decision maker was indifferent between a guaranteed payoff of M and a lottery in which the best payoff is obtained with probability p and the worst payoff is obtained with probability $(1 - p)$. Thus, the utility for any monetary value would have been equal to the probability of earning the best payoff. Often this choice is made because of the ease in computation. We chose not to do so to emphasize the distinction between the utilities and the indifference probabilities for the lottery.
2. Circumstances often dictate whether one acts as a risk avoider or a risk taker when making a decision. For example, you may think of yourself as a risk avoider when faced with financial decisions, but if you have ever purchased a lottery ticket, you have actually acted as a risk taker. Suppose you purchase a \$1 lottery ticket for a simple lottery in which the object is to pick the six numbers that will be drawn from 50 potential numbers. Also suppose that the winner (who correctly chooses all six numbers that are drawn) will receive \$1,000,000. There are 15,890,700 possible winning combinations, so your probability of winning is 1/15,890,700 = 0.00000062929889809763 (i.e., *very low*) and the expected value of your ticket is

$$\frac{1}{15{,}890{,}700}(\$1{,}000{,}000 - \$1) + \left(1 - \frac{1}{15{,}890{,}700}\right)(-\$1) = -\$0.93707$$

or about −\$0.94.

If a lottery ticket has a negative expected value, why does anyone play? The answer is in utility. Many people associate great utility with the possibility of winning the \$1,000,000 prize and relatively little utility with the \$1 cost for a ticket, and so the expected value of the utility of the lottery ticket is positive even though the expected value of the ticket is negative.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary

Decision analysis can be used to determine a recommended decision alternative or an optimal decision strategy when a decision maker is faced with an uncertain and risk-filled pattern of future events. The goal of decision analysis is to identify the best decision alternative or the optimal decision strategy, given information about the uncertain events and the possible consequences or payoffs. The "best" decision should consider the risk preference of the decision maker in evaluating outcomes.

We showed how payoff tables and decision trees could be used to structure a decision problem and describe the relationships among the decisions, the chance events, and the consequences. We presented three approaches to decision making without probabilities: the optimistic approach, the conservative approach, and the minimax regret approach. When probability assessments are provided for the states of nature, the expected value approach can be used to identify the recommended decision alternative or decision strategy.

Even though the expected value approach can be used to obtain a recommended decision alternative or optimal decision strategy, the payoff that actually occurs will usually have a value different from the expected value. A risk profile provides a probability distribution for the possible payoffs and can assist the decision maker in assessing the risks associated with different decision alternatives. Sensitivity analysis can be conducted to determine the effect changes in the probabilities for the states of nature and changes in the values of the payoffs have on the recommended decision alternative.

In cases in which sample information about the chance events is available, a sequence of decisions has to be made. First we must decide whether to obtain the sample information. If the answer is yes, an optimal decision strategy based on the specific sample information must be developed. In this situation, decision trees and the expected value approach can be used to determine the optimal decision strategy.

Bayes' theorem can be used to compute branch probabilities for decision trees. Bayes' theorem updates a decision maker's prior probabilities regarding the states of nature using sample information to compute revised posterior probabilities.

We showed how utility could be used in decision-making situations in which monetary value did not provide an adequate measure of the payoffs. Utility is a measure of the total worth of an outcome. As such, utility takes into account the decision maker's assessment of all aspects of a consequence, including profit, loss, risk, and perhaps additional nonmonetary factors. The examples showed how the use of expected utility can lead to decision recommendations that differ from those based on expected value.

A decision maker's judgment must be used to establish the utility for each consequence. We presented a step-by-step procedure to determine a decision maker's utility for monetary payoffs. We also discussed how conservative, risk-avoiding decision makers assess utility differently from more aggressive, risk-taking decision makers.

Glossary

Bayes' theorem A theorem that enables the use of sample information to revise prior probabilities.

Branch Lines showing the alternatives from decision nodes and the outcomes from chance nodes.

Chance event An uncertain future event affecting the consequence, or payoff, associated with a decision.

Chance nodes Nodes indicating points at which an uncertain event will occur.

Conditional probabilities The probability of one event, given the known outcome of a (possibly) related event.

Conservative approach An approach to choosing a decision alternative without using probabilities. For a maximization problem, it leads to choosing the decision alternative that maximizes the minimum payoff; for a minimization problem, it leads to choosing the decision alternative that minimizes the maximum payoff.

Decision alternatives Options available to the decision maker.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Decision nodes Nodes indicating points at which a decision is made.
Decision strategy A strategy involving a sequence of decisions and chance outcomes to provide the optimal solution to a decision problem.
Decision tree A graphical representation of the decision problem that shows the sequential nature of the decision-making process.
Expected utility (EU) The weighted average of the utilities associated with a decision alternative. The weights are the state-of-nature probabilities.
Expected value (EV) For a chance node, the weighted average of the payoffs. The weights are the state-of-nature probabilities.
Expected value approach An approach to choosing a decision alternative based on the expected value of each decision alternative. The recommended decision alternative is the one that provides the best expected value.
Expected value of perfect information (EVPI) The difference between the expected value of an optimal strategy based on perfect information and the "best" expected value without any sample information.
Expected value of sample information (EVSI) The difference between the expected value of an optimal strategy based on sample information and the "best" expected value without any sample information.
Minimax regret approach An approach to choosing a decision alternative without using probabilities. For each alternative, the maximum regret is computed, which leads to choosing the decision alternative that minimizes the maximum regret.
Node An intersection or junction point of a decision tree.
Optimistic approach An approach to choosing a decision alternative without using probabilities. For a maximization problem, it leads to choosing the decision alternative corresponding to the largest payoff; for a minimization problem, it leads to choosing the decision alternative corresponding to the smallest payoff.
Outcome The result obtained when a decision alternative is chosen and a chance event occurs.
Payoff A measure of the outcome of a decision such as profit, cost, or time. Each combination of a decision alternative and a state of nature has an associated payoff.
Payoff table A tabular representation of the payoffs for a decision problem.
Perfect information A special case of sample information in which the information tells the decision maker exactly which state of nature is going to occur.
Posterior (revised) probabilities The probabilities of the states of nature after revising the prior probabilities based on sample information.
Prior probabilities The probabilities of the states of nature prior to obtaining sample information.
Regret (opportunity loss) The amount of loss (lower profit or higher cost) from not making the best decision for each state of nature.
Risk analysis The study of the possible payoffs and probabilities associated with a decision alternative or a decision strategy in the face of uncertainty.
Risk avoider A decision maker who would choose a guaranteed payoff over a lottery with a better expected payoff.
Risk neutral A decision maker for whom the decision alternative with the best expected value is identical to the alternative with the highest expected utility.
Risk profile The probability distribution of the possible payoffs associated with a decision alternative or decision strategy.
Risk taker A decision maker who would choose a lottery over a better guaranteed payoff.
Sample information New information obtained through research or experimentation that enables updating or revising the state-of-nature probabilities.
Sensitivity analysis The study of how changes in the probability assessments for the states of nature or changes in the payoffs affect the recommended decision alternative.
States of nature The possible outcomes for chance events that affect the payoff associated with a decision alternative.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Utility A measure of the total worth of a consequence reflecting a decision maker's attitude toward considerations such as profit, loss, and risk.
Utility function for money A curve that depicts the relationship between monetary value and utility.

Problems

1. **Two Decision Alternatives and Three States of Nature.** The following payoff table shows profit for a decision analysis problem with two decision alternatives and three states of nature. LO 1, 2

	State of Nature		
Decision Alternative	s_1	s_2	s_3
d_1	250	100	25
d_2	100	100	75

   a. Construct a decision tree for this problem.
   b. If the decision maker knows nothing about the probabilities of the three states of nature, what is the recommended decision using the optimistic, conservative, and minimax regret approaches?

2. **Plant Size Decision.** Southland Corporation's decision to produce a new line of recreational products resulted in the need to construct either a small plant or a large plant. The best selection of plant size depends on how the marketplace reacts to the new product line. To conduct an analysis, marketing management has decided to view the possible long-run demand as low, medium, or high. The following payoff table shows the projected profit in millions of dollars. LO 1, 2

	Long-Run Demand		
Plan Size	Low	Medium	High
Small	150	200	200
Large	50	200	500

   a. What is the decision to be made, and what is the chance event for Southland's problem?
   b. Construct a decision tree.
   c. Recommend a decision based on the use of the optimistic, conservative, and minimax regret approaches.

DATA *file* carleases

3. **Car Leases.** Amy Lloyd is interested in leasing a new Honda and has contacted three automobile dealers for pricing information. Each dealer offered Amy a closed-end 36-month lease with no down payment due at the time of signing. Each lease includes a monthly charge and a mileage allowance. Additional miles receive a surcharge on a per-mile basis. The monthly lease cost, the mileage allowance, and the cost for additional miles are as follows:

Dealer	Monthly Cost ($)	Mileage Allowance	Cost per Additional Mile ($)
Hepburn Honda	299	36,000	0.15
Midtown Motors	310	45,000	0.20
Hopkins Automotive	325	54,000	0.15

Amy decided to choose the lease option that will minimize the total 36-month cost. However, Amy is uncertain about how many miles will be driven in this car over the next three years. For purposes of this decision, Amy believes it is reasonable to

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

assume that the car will be driven 12,000 miles per year, 15,000 miles per year, or 18,000 miles per year. With this assumption, Amy estimated the total costs for the three lease options. For example, Amy figures that the Hepburn Honda lease will cost $36(\$299) + \$0.15(36{,}000 - 36{,}000) = \$10{,}764$ if the car is driven 12,000 miles per year, $36(\$299) + \$0.15(45{,}000 - 36{,}000) = \$12{,}114$ if the car is driven 15,000 miles per year, or $36(\$299) + \$0.15(54{,}000 - 36{,}000) = \$13{,}464$ if the car is driven 18,000 miles per year. **LO 1, 2, 3, 4**

a. What is the decision, and what is the chance event?
b. Construct a payoff table for Amy's problem.
c. If Amy has no idea which of the three mileage assumptions is most appropriate, what is the recommended decision (leasing option) using the optimistic, conservative, and minimax regret approaches?
d. Suppose that the probabilities that Amy drives 12,000, 15,000, and 18,000 miles per year are 0.5, 0.4, and 0.1, respectively. What option should Amy choose using the expected value approach?
e. Develop a risk profile for the decision selected in part (d). What is the most likely cost, and what is its probability?
f. Suppose that, after further consideration, Amy concludes that the probabilities that the car will be driven 12,000, 15,000, and 18,000 miles per year are 0.3, 0.4, and 0.3, respectively. What decision should Amy make using the expected value approach?

4. **Market Segment Investment.** Investment advisors estimated the stock market returns for four market segments: technology, financial, manufacturing, and pharmaceuticals. Annual return projections vary depending on whether the general economic conditions are improving, stable, or declining. The anticipated annual return percentages for each market segment under each economic condition are shown in the following table. **LO 3**

DATA*file* market_investment

	Economic Condition		
Market Segment	**Improving**	**Stable**	**Declining**
Technology	10	2	−4
Financial	8	5	−3
Manufacturing	6	4	−2
Pharmaceuticals	6	5	−1

a. Assume that an individual investor wants to select one market segment for a new investment. A forecast shows improving to declining economic conditions with the following probabilities: improving (0.2), stable (0.5), and declining (0.3). What is the preferred market segment for the investor, and what is the expected return percentage?
b. At a later date, a revised forecast shows a potential for an improvement in economic conditions. New probabilities are as follows: improving (0.4), stable (0.4), and declining (0.2). What is the preferred market segment for the investor based on these new probabilities? What is the expected return percentage?

5. **Data Warehouse Operation.** Hudson Corporation is considering three options for managing its data warehouse: continuing with its own staff, hiring an outside vendor to do the managing, or using a combination of its own staff and an outside vendor. The cost of the operation depends on future demand. The annual cost of each option (in thousands of dollars) depends on demand as shown in the following table. **LO 3, 4**

DATA*file* datawarehouse

	Demand		
Staffing Options	**High**	**Medium**	**Low**
Own staff	650	650	600
Outside vendor	900	600	300
Combination	800	650	500

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

a. If the demand probabilities are 0.2, 0.5, and 0.3, which decision alternative will minimize the expected cost of the data warehouse? What is the expected annual cost associated with that recommendation?
b. Construct a risk profile for the optimal decision in part (a). What is the probability of the cost exceeding $700,000?

6. **Two States of Nature and Two Decision Alternatives.** The following payoff table shows the profit for a decision problem with two states of nature and two decision alternatives. **LO 3, 5**

	State of Nature	
Decision Alternative	s_1	s_2
d_1	10	1
d_2	4	3

a. Suppose $P(s_1) = 0.2$ and $P(s_2) = 0.8$. What is the best decision using the expected value approach?
b. Perform sensitivity analysis on the payoffs for decision alternative d_1. Assume that the probabilities are as given in part (a), and find the range of payoffs under states of nature s_1 and s_2 that will keep the solution found in part (a) optimal. Is the solution more sensitive to the payoff under state of nature s_1 or s_2?

7. **Cleveland to Myrtle Beach Air Service.** Myrtle Air Express decided to offer direct service from Cleveland to Myrtle Beach. Management must decide between a full-price service using the company's new fleet of jet aircraft and a discount service using smaller-capacity commuter planes. It is clear that the best choice depends on the market reaction to the service Myrtle Air offers. Management developed estimates of the contribution to profit for each type of service based on two possible levels of demand for service to Myrtle Beach: strong and weak. The following table shows the estimated quarterly profits (in thousands of dollars). **LO 1, 2, 3, 5**

	Demand for Service	
Service	**Strong**	**Weak**
Full price	$960	−$490
Discount	$670	$320

a. What is the decision to be made, what is the chance event, and what is the consequence for this problem? How many decision alternatives are there? How many outcomes are there for the chance event?
b. If nothing is known about the probabilities of the chance outcomes, what is the recommended decision using the optimistic, conservative, and minimax regret approaches?
c. Suppose that management of Myrtle Air Express believes that the probability of strong demand is 0.7 and the probability of weak demand is 0.3. Use the expected value approach to determine an optimal decision.
d. Suppose that the probability of strong demand is 0.8 and the probability of weak demand is 0.2. What is the optimal decision using the expected value approach?
e. Use sensitivity analysis to determine the range of demand probabilities for which each of the decision alternatives has the largest expected value.

8. **Video Game Profitability.** Video Tech is considering marketing one of two new video games for the coming holiday season: Battle Pacific or Space Pirates. Battle Pacific

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

is a unique game and appears to have no competition. Estimated profits (in thousands of dollars) under high, medium, and low demand are as follows:

	Demand		
Battle Pacific	**High**	**Medium**	**Low**
Profit	$1,000	$700	$300
Probability	0.2	0.5	0.3

Video Tech is optimistic about its Space Pirates game. However, the concern is that profitability will be affected by a competitor's introduction of a video game viewed as similar to Space Pirates. Estimated profits (in thousands of dollars) with and without competition are shown in the following tables. **LO 1, 3, 5**

DATA*file* spacepirates

Space Pirates		Demand	
With Competition	**High**	**Medium**	**Low**
Profit	$800	$400	$200
Probability	0.3	0.4	0.3

Space Pirates		Demand	
Without Competition	**High**	**Medium**	**Low**
Profit	$1,600	$800	$400
Probability	0.5	0.3	0.2

a. Develop a decision tree for the Video Tech problem.
b. For planning purposes, Video Tech believes there is a 0.6 probability that its competitor will produce a new game similar to Space Pirates. Given this probability of competition, the director of planning recommends marketing the Battle Pacific video game. Using expected value, what is your recommended decision?
c. Show a risk profile for your recommended decision.
d. Use sensitivity analysis to determine what the probability of competition for Space Pirates would have to be for you to change your recommended decision alternative.

9. **Augmented Reality Application Investment.** Software applications that use augmented reality can provide an enhanced version of the real-world by adding digital elements such as visual graphics, sounds, and other sensory feedback to the user. Augmented reality has been applied to settings such as healthcare, military, and entertainment. According to a 2022 article by *Yahoo! Finance*, the global market size of augmented reality software applications is estimated to surpass $332 billion in 2028. However, there is considerable uncertainty in this estimated market size because the technologies to support augmented reality are still being developed and there is little history on its use.

Consider the case of AR Ventures, a venture-capital firm that provides support for start-up companies using augmented reality technology. AR Ventures has the opportunity to invest in one of two start-ups utilizing augmented reality technology: one that is developing a healthcare application and one that is developing an entertainment application. The healthcare application uses augmented reality to help train medical students, and the entertainment application allows people to compete virtually in the sport of badminton. AR Ventures believes that the potential market size of the healthcare application is larger, but that the technology needed for this application is unproven.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The technology used for the entertainment application is more well established, but the market size is likely smaller.

Suppose that demand for the entertainment application is estimated to be high, medium, or low with the following associated revenues, costs, and probabilities.

DATA*file*
ar_ventures

	Demand		
Entertainment Application	**High**	**Medium**	**Low**
Revenue ($ millions)	1.80	1.00	0.25
Cost ($ millions)	1.10	0.80	0.60
Probability	0.40	0.30	0.30

The costs associated with the healthcare application are largely dependent on the success of one particular new device that would be worn by the medical students. If the current protype of the device proves to be successful, the costs will be significantly lower than if the prototype device is not successful. The estimated revenues, costs, and probabilities of different demand scenarios for each possibility follow. AR Ventures would like to make the decision of which application in which to invest based on profit (revenue − cost). **LO 1, 3**

	Demand		
Healthcare Application (If Prototype Is Successful)	**High**	**Medium**	**Low**
Revenue ($ millions)	6.50	4.00	2.50
Cost ($ millions)	4.30	3.40	2.80
Probability	0.30	0.50	0.20

	Demand		
Healthcare Application (If Prototype Is Unsuccessful)	**High**	**Medium**	**Low**
Revenue ($ millions)	6.50	4.00	2.50
Cost ($ millions)	5.90	4.10	3.60
Probability	0.30	0.50	0.20

a. Develop a decision tree for the AR Ventures problem.
b. Suppose that AR Ventures believes it is equally likely that the healthcare application device prototype will be successful as it is that the prototype will be unsuccessful. Using an expected value approach would you recommend that AR Ventures invests in the entertainment application or the healthcare application?
c. Would your recommendation in part (b) change if the probability of the prototype being successful is only 0.25? Why or why not?

10. **Chardonnay or Riesling Grapes.** Seneca Hill Winery recently purchased land for the purpose of establishing a new vineyard. Management is considering two varieties of white grapes for the new vineyard: Chardonnay and Riesling. The Chardonnay grapes would be used to produce a dry Chardonnay wine, and the Riesling grapes would be used to produce a semidry Riesling wine. It takes approximately four years from the time of planting before new grapes can be harvested. This length of time creates a great deal of uncertainty concerning future demand and makes the decision about the type of grapes to plant difficult. Three possibilities are being considered: Chardonnay grapes only; Riesling grapes only; and both Chardonnay and Riesling grapes. Seneca management decided that for planning purposes it would be adequate to consider only

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

two demand possibilities for each type of wine: strong or weak. With two possibilities for each type of wine, it was necessary to assess four probabilities. With the help of some forecasts in industry publications, management made the following probability assessments:

DATA*file* winery

	Riesling Demand (Probabilities)	
Chardonnay Demand	**Weak**	**Strong**
Weak	0.05	0.50
Strong	0.25	0.20

Revenue projections show an annual contribution to profit of $20,000 if Seneca Hill plants only Chardonnay grapes and demand is weak for Chardonnay wine, and $70,000 if Seneca plants only Chardonnay grapes and demand is strong for Chardonnay wine. If Seneca plants only Riesling grapes, the annual profit projection is $25,000 if demand is weak for Riesling grapes and $45,000 if demand is strong for Riesling grapes. If Seneca plants both types of grapes, the annual profit projections are as shown in the table that follows. **LO 1, 3**

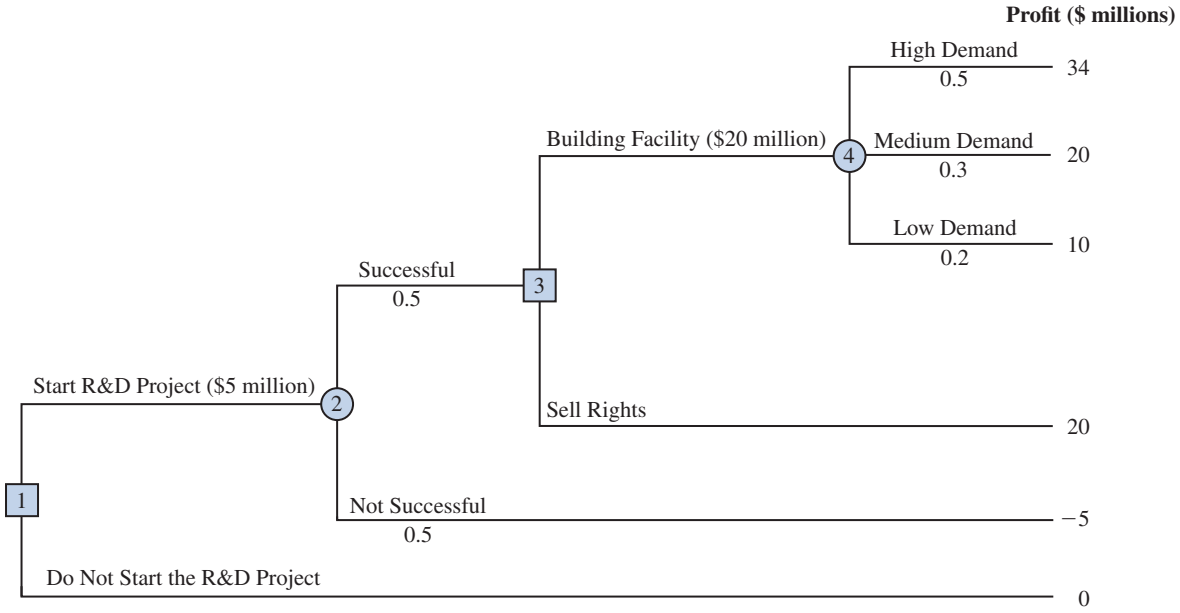
	Riesling Demand (Profits)	
Chardonnay Demand	**Weak**	**Strong**
Weak	$22,000	$40,000
Strong	$26,000	$60,000

a. What is the decision to be made, what is the chance event, and what is the consequence? Identify the alternatives for the decisions and the possible outcomes for the chance events.
b. Develop a decision tree.
c. Use the expected value approach to recommend which alternative Seneca Hill Winery should follow in order to maximize expected annual profit.
d. Suppose management is concerned about the probability assessments when demand for Chardonnay wine is strong. Some believe it is likely for Riesling demand to also be strong in this case. Suppose that the probability of strong demand for Chardonnay and weak demand for Riesling is 0.05 and that the probability of strong demand for Chardonnay and strong demand for Riesling is 0.40. How does this change the recommended decision? Assume that the probabilities when Chardonnay demand is weak are still 0.05 and 0.50.
e. Other members of the management team expect the Chardonnay market to become saturated at some point in the future, causing a fall in prices. Suppose that the annual profit projections fall to $50,000 when demand for Chardonnay is strong and only Chardonnay grapes are planted. Using the original probability assessments, determine how this change would affect the optimal decision.

11. **R&D Project.** Hemmingway, Inc. is considering a $5 million research and development (R&D) project. Profit projections appear promising, but Hemmingway's president is concerned because the probability that the R&D project will be successful is only 0.50. Furthermore, the president knows that even if the project is successful, it will require that the company build a new production facility at a cost of $20 million in order to manufacture the product. If the facility is built, uncertainty remains about the demand and thus uncertainty about the profit that will be realized. Another option is that if the R&D project is successful, the company could sell the rights to the product for an estimated $25 million. Under this option, the company would not build the $20 million production facility.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

The decision tree follows. The profit projection for each outcome is shown at the end of the branches. For example, the revenue projection for the high demand outcome is \$59 million. However, the cost of the R&D project (\$5 million) and the cost of the production facility (\$20 million) show the profit of this outcome to be \$59 − \$5 − \$20 = \$34 million. Branch probabilities are also shown for the chance events. **LO 1, 3, 4**



a. Analyze the decision tree to determine whether the company should undertake the R&D project. If it does, and if the R&D project is successful, what should the company do? What is the expected value of your strategy?
b. What must the selling price be for the company to consider selling the rights to the product?
c. Develop a risk profile for the optimal strategy.

12. **Medical Decision Making for Cancer Patients.** Cancer patients and their physicians face many decisions related to treatment options. Depending on the type of cancer, the location of the cancer, and patient history, treatment options can include surgery, chemotherapy, radiation therapy, and immunotherapy. The probability of the success of these treatments is affected by many factors based on the type of cancer and overall health of the patient. Medical decision makers often measure the expected effectiveness of different types of treatments using a metric called quality adjusted life year (QALY). One QALY is equivalent to one year of life in perfect health. Decision makers use QALYs to approximate the possible outcomes on the patient from different treatment options.

Suppose that a physician is working with a cancer patient to determine the best course of treatment. The physician and the patient need to decide whether to have an exploratory surgery procedure to determine the spread of the cancer or to follow an active surveillance strategy in which no surgery is performed but the patient has regular tests to track progression of the cancer. If exploratory surgery is chosen, the results of the surgery can be favorable (F), unfavorable (U), or show that no further treatment is needed (N). Past data suggests that $P(F) = 0.5$, $P(U) = 0.2$, and $P(N) = 0.3$. If the exploratory surgery is favorable or unfavorable, then one of two treatments will be used, Treatment A (A) or Treatment B (B) using the following conditional probabilities:

$$P(A|F) = 0.4 \qquad P(A|U) = 0.8$$
$$P(B|F) = 0.6 \qquad P(B|U) = 0.2$$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Possible patient outcomes are measured using QALYs, and this depends on both the result of the exploratory surgery and the treatment option required as shown in the following table.

DATA*file* cancerpatients

Outcome of Exploratory Surgery	Treatment Option Required	Resulting QALYs
Favorable	A	22
Favorable	B	26
Unfavorable	A	14
Unfavorable	B	17
No Treatment Needed	None	35

If active surveillance is chosen, then the cancer can either progress (P) or not progress (NP) where $P(P) = 0.4$ and $P(NP) = 0.6$. If the cancer does not progress, then no further treatment is needed. If the cancer does progress, then either Treatment A or Treatment B will be required where $P(A|P) = 0.7$ and $P(B|P) = 0.3$. The possible resulting QALYs for this decision alternative follow. **LO 1, 3**

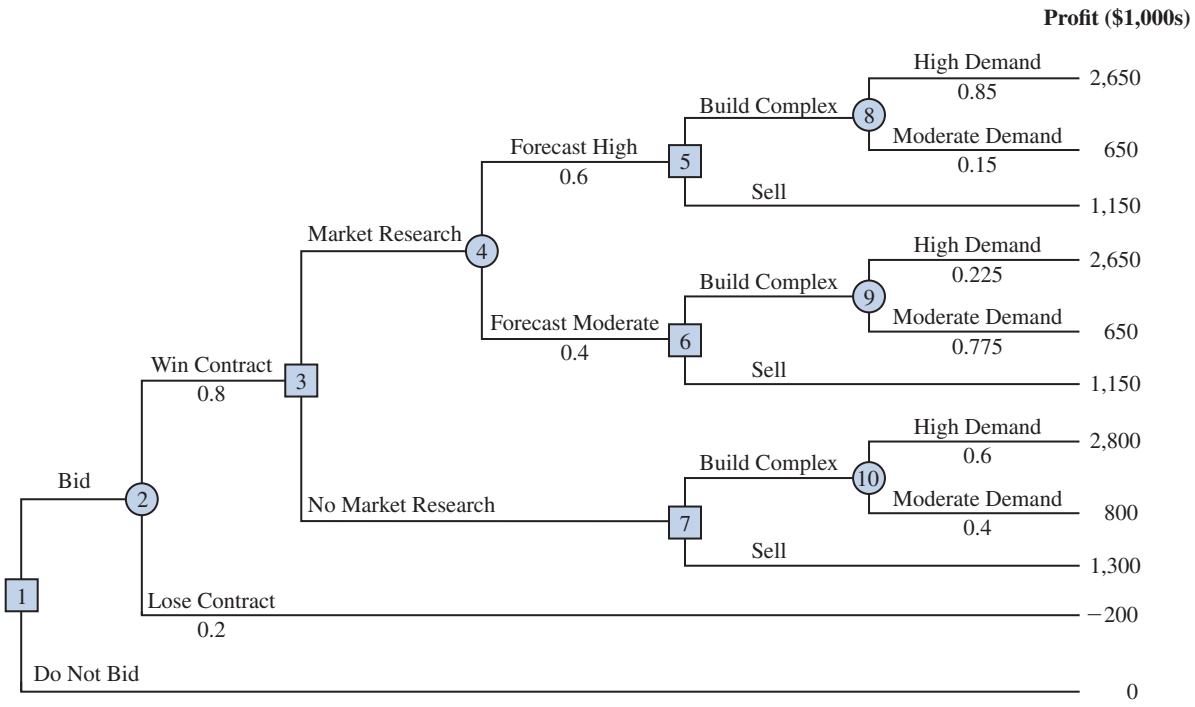
Outcome of Active Surveillance	Treatment Option Required	Resulting QALYs
Progress	A	15
Progress	B	20
Not Progress	None	35

a. Construct a decision tree for this problem.
b. Using the expected value approach and QALYs as the outcome measure, would you recommend exploratory surgery or active surveillance? What is the expected QALYs associated with this decision?

13. **New Office Building Complex.** Dante Development Corporation is considering bidding on a contract for a new office building complex. The following figure shows the decision tree prepared by one of Dante's analysts. At node 1, the company must decide whether to bid on the contract. The cost of preparing the bid is $200,000. The upper branch from node 2 shows that the company has a 0.8 probability of winning the contract if it submits a bid. If the company wins the bid, it will have to pay $2 million to become a partner in the project. Node 3 shows that the company will then consider doing a market research study to forecast demand for the office units prior to beginning construction. The cost of this study is $150,000. Node 4 is a chance node showing the possible outcomes of the market research study.

Nodes 5, 6, and 7 are similar in that they are the decision nodes for Dante to either build the office complex or sell the rights in the project to another developer. The decision to build the complex will result in an income of $5 million if demand is high and $3 million if demand is moderate. If Dante chooses to sell its rights in the project to another developer, income from the sale is estimated to be $3.5 million. The probabilities shown at nodes 4, 8, and 9 are based on the projected outcomes of the market research study. **LO 1, 3, 4, 6**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.



a. Verify Dante's profit projections shown at the ending branches of the decision tree by calculating the payoffs of $2,650,000 and $650,000 for first two outcomes.
b. What is the optimal decision strategy for Dante, and what is the expected profit for this project?
c. What would the cost of the market research study have to be before Dante would change its decision about the market research study?
d. Develop a risk profile for Dante.

14. **New College Textbook.** Embassy Publishing Company received a six-chapter manuscript for a new college textbook. The editor of the college division is familiar with the manuscript and estimated a 0.65 probability that the textbook will be successful. If successful, a profit of $750,000 will be realized. If the company decides to publish the textbook and it is unsuccessful, a loss of $250,000 will occur.

Before making the decision to accept or reject the manuscript, the editor is considering sending the manuscript out for review. A review process provides either a favorable (F) or unfavorable (U) evaluation of the manuscript. Past experience with the review process suggests that probabilities $P(F) = 0.7$ and $P(U) = 0.3$ apply. Let s_1 = the textbook is successful and s_2 = the textbook is unsuccessful. The editor's initial probabilities of s_1 and s_2 will be revised based on whether the review is favorable or unfavorable. The revised probabilities follow. **LO 1, 3, 6**

$$P(s_1|F) = 0.85 \qquad P(s_1|U) = 0.417$$
$$P(s_2|F) = 0.15 \qquad P(s_2|U) = 0.583$$

a. Construct a decision tree assuming that the company will first make the decision as to whether to send the manuscript out for review and then make the decision to accept or reject the manuscript.
b. Analyze the decision tree to determine the optimal decision strategy for the publishing company.
c. If the manuscript review costs $15,000, what is your recommendation?
d. What is the expected value of perfect information? What does this EVPI suggest for the company?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

15. **Value of Perfect Information.** The following profit payoff table was presented in Problem 1:

Decision Alternative	State of Nature s_1	s_2	s_3
d_1	250	100	25
d_2	100	100	75

The probabilities for the states of nature are $P(s_1) = 0.65$, $P(s_2) = 0.15$, and $P(s_3) = 0.20$. **LO 1, 3, 6**

a. What is the optimal decision strategy if perfect information were available?
b. What is the expected value for the decision strategy developed in part (a)?
c. Using the expected value approach, what is the recommended decision without perfect information? What is its expected value?
d. What is the expected value of perfect information?

16. **Lake Placid Community Center.** The Lake Placid Town Council decided to build a new community center to be used for conventions, concerts, and other public events, but considerable controversy surrounds the appropriate size. Many influential citizens want a large center that would be a showcase for the area. But the mayor feels that if demand does not support such a center, the community will lose a large amount of money. To provide structure for the decision process, the council narrowed the building alternatives to three sizes: small, medium, and large. Everybody agreed that the critical factor in choosing the best size is the number of people who will want to use the new facility. A regional planning consultant provided demand estimates under three scenarios: worst case, base case, and best case. The worst-case scenario corresponds to a situation in which tourism drops substantially; the base-case scenario corresponds to a situation in which Lake Placid continues to attract visitors at current levels; and the best-case scenario corresponds to a substantial increase in tourism. The consultant has provided probability assessments of 0.10, 0.60, and 0.30 for the worst-case, base-case, and best-case scenarios, respectively.

The town council suggested using net cash flow over a five-year planning horizon as the criterion for deciding on the best size. The following projections of net cash flow (in thousands of dollars) for a five-year planning horizon have been developed. All costs, including the consultant's fee, have been included. **LO 1, 3, 4, 6**

DATA*file* lakeplacid

Center Size	Demand Scenario Worst Case	Base Case	Best Case
Small	400	500	660
Medium	−250	650	800
Large	−400	580	990

a. What decision should Lake Placid make using the expected value approach?
b. Construct risk profiles for the medium and large alternatives. Given the mayor's concern over the possibility of losing money and the result of part (a), which alternative would you recommend?
c. Compute the expected value of perfect information. Do you think it would be worth trying to obtain additional information concerning which scenario is likely to occur?
d. Suppose the probability of the worst-case scenario increases to 0.2, the probability of the base-case scenario decreases to 0.5, and the probability of the best-case scenario remains at 0.3. What effect, if any, would these changes have on the decision recommendation?

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

e. The consultant has suggested that an expenditure of $150,000 on a promotional campaign over the planning horizon will effectively reduce the probability of the worst-case scenario to zero. If the campaign can be expected to also increase the probability of the best-case scenario to 0.4, is it a good investment?

17. **Rezoning Property.** A real estate investor has the opportunity to purchase land currently zoned as residential. If the county board approves a request to rezone the property as commercial within the next year, the investor will be able to lease the land to a large discount firm that wants to open a new store on the property. However, if the zoning change is not approved, the investor will have to sell the property at a loss. Profits (in thousands of dollars) are shown in the payoff table that follows. **LO 1, 3, 6**

	State of Nature	
Decision Alternative	**Rezoning Approved, s_1**	**Rezoning Not Approved, s_2**
Purchase, d_1	600	−200
Do not purchase, d_2	0	0

a. If the probability that the rezoning will be approved is 0.5, what decision is recommended? What is the expected profit?

b. The investor can purchase an option to buy the land. Under the option, the investor maintains the rights to purchase the land anytime during the next three months while learning more about possible resistance to the rezoning proposal from area residents. Probabilities are as follows:

$$\text{Let} \quad H = \text{high resistance to rezoning}$$
$$L = \text{low resistance to rezoning}$$

$$P(H) = 0.55 \quad P(s_1|H) = 0.18 \quad P(s_2|H) = 0.82$$
$$P(L) = 0.45 \quad P(s_1|L) = 0.89 \quad P(s_2|L) = 0.11$$

What is the optimal decision strategy if the investor uses the option period to learn more about the resistance from area residents before making the purchase decision?

c. If the option will cost the investor an additional $10,000, should the investor purchase the option? Why or why not? What is the maximum that the investor should be willing to pay for the option?

18. **Posterior Probabilities Calculation.** Suppose that you are given a decision situation with three possible states of nature: s_1, s_2, and s_3. The prior probabilities are $P(s_1) = 0.2$, $P(s_2) = 0.5$, and $P(s_3) = 0.3$. With sample information I, $P(I|s_1) = 0.1$, $P(I|s_2) = 0.05$, and $P(I|s_3) = 0.2$. Compute the revised (or posterior) probabilities: $P(s_1|I)$, $P(s_2|I)$, and $P(s_3|I)$. **LO 7**

19. **Carpool Route.** To save on expenses, Rona and Jerry agreed to form a carpool for traveling to and from work. Rona prefers to use the somewhat longer but more consistent Queen City Avenue. Although Jerry prefers the quicker expressway, Jerry agreed with Rona that they should take Queen City Avenue if the expressway has a traffic jam. The following payoff table provides the one-way time estimate in minutes for traveling to or from work:

	State of Nature	
Decision Alternative	**Expressway Open, s_1**	**Expressway Jammed, s_2**
Queen City Avenue, d_1	30	30
Expressway, d_2	25	45

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Based on their experience with traffic problems, Rona and Jerry agreed on a 0.15 probability that the expressway would be jammed.

In addition, they agreed that weather seemed to affect the traffic conditions on the expressway. Let

C = clear
O = overcast
R = rain

The following conditional probabilities apply:

$$P(C|s_1) = 0.8 \quad P(O|s_1) = 0.2 \quad P(R|s_1) = 0.0$$
$$P(C|s_2) = 0.1 \quad P(O|s_2) = 0.3 \quad P(R|s_2) = 0.6$$

Use this information to complete the following. **LO 1, 3, 7**

a. Use Bayes' theorem for probability revision to compute the probability of each weather condition and the conditional probability of the expressway being open, s_1, or jammed, s_2, given each weather condition.
b. Show the decision tree for this problem.
c. What is the optimal decision strategy, and what is the expected travel time?

20. **Manufacture or Purchase.** The Gorman Manufacturing Company must decide whether to manufacture a component part at its Milan, Michigan, plant or purchase the component part from a supplier. The resulting profit is dependent on the demand for the product. The following payoff table shows the projected profit:

	State of Nature		
Decision Alternative	Low Demand s_1	Medium Demand s_2	High Demand s_3
Manufacture, d_1	−$20,000	$40,000	$100,000
Purchase, d_2	$10,000	$45,000	$70,000

The state-of-nature probabilities are $P(s_1) = 0.35$, $P(s_2) = 0.35$, and $P(s_3) = 0.30$. **LO 1, 3, 6**

a. Use a decision tree to recommend a decision.
b. Use EVPI to determine whether Gorman should attempt to obtain a better estimate of demand.
c. A test market study of the potential demand for the product is expected to report either a favorable (F) or an unfavorable (U) condition. The relevant conditional probabilities are as follows:

$$P(F|s_1) = 0.10 \; P(U|s_1) = 0.90$$
$$P(F|s_2) = 0.40 \; P(U|s_2) = 0.60$$
$$P(F|s_3) = 0.60 \; P(U|s_3) = 0.40$$

What is the probability that the market research report will be favorable? *Hint:* We can find this value by summing the joint probability values as follows: $P(F) = P(F \cap s_1) + P(F \cap s_2) + P(F \cap s_3) = P(s_1)P(F|s_1) + P(s_2)P(F|s_2) + P(s_3)P(F|s_3)$.

Joint probabilities are discussed in Chapter 5.

d. What is Gorman's optimal decision strategy?
e. What is the expected value of the market research information?

21. **Calculating Utility.** The following table displays the possible payoffs from a particular investment and the indifference probabilities for an investor. **LO 8, 9**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA*file*
indifference_prob

Payoff ($)	Indifference Probability
300,000	Does not apply
250,000	0.70
200,000	0.60
125,000	0.40
50,000	0.30
0	0.25
−25,000	0.20
−100,000	0.10
−225,000	Does not apply

a. Assume that the lowest payoff is assigned a utility of 0 and that the greatest payoff is assigned a utility of 10. Use these values to calculate utilities for each of the possible payoff values.
b. Graph the utility function for this investor. Would you classify the investor as a risk avoider, a risk taker, or risk neutral?

22. **Investment Alternatives.** A firm has three investment alternatives. Payoffs are in thousands of dollars. **LO 1, 3, 8, 9**

DATA*file*
investment_alternatives

	Economic Conditions		
Decision Alternative	Up, s_1	Stable, s_2	Down, s_3
Investment A, d_1	100	25	0
Investment B, d_2	75	50	25
Investment C, d_3	50	50	50
Probabilities	0.40	0.30	0.30

a. Using the expected value approach, which decision is preferred?
b. For the lottery having a payoff of $100,000 with probability p and $0 with probability $(1 - p)$, two decision makers expressed the following indifference probabilities. Find the most preferred decision for each decision maker using the expected utility approach.

	Indifference Probability (p)	
Profit ($)	Decision Maker A	Decision Maker B
75,000	0.80	0.60
50,000	0.60	0.30
25,000	0.30	0.15

c. Why don't decision makers A and B select the same decision alternative?

23. **Insurance Policy for New Office Building.** Alexander Industries is considering purchasing an insurance policy for its new office building in St. Louis, Missouri. The policy has an annual cost of $10,000. If Alexander Industries doesn't purchase the insurance and minor fire damage occurs, a cost of $100,000 is anticipated; the cost if major or total destruction occurs is $200,000. The costs, including the state-of-nature probabilities, are shown in the following table. **LO 1, 3, 8**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

DATA *file* office_insurance

Decision Alternative	Damage: None, s_1	Minor, s_2	Major, s_3
Purchase insurance, d_1	$10,000	$10,000	$10,000
Do not purchase insurance, d_2	$0	$100,000	$200,000
Probabilities	0.96	0.03	0.01

a. Using the expected value approach, what decision do you recommend?
b. What lottery would you use to assess utilities? *Note:* Because the data are costs, the best payoff is $0.
c. Assume that you found the following indifference probabilities for the lottery defined in part (b). What decision would you recommend?

Cost ($)	Indifference Probability
10,000	$p = 0.99$
100,000	$p = 0.60$

d. Do you favor using expected value or expected utility for this decision problem? Why?

24. **Lottery Ticket.** In a certain state lottery, a lottery ticket costs $2. In terms of the decision to purchase or not to purchase a lottery ticket, assume that the payoff table that follows applies. **LO 1, 3, 8**

Decision Alternatives	State of Nature: Win, s_1	Lose, s_2
Purchase lottery ticket, d_1	$300,000	−$2
Do not purchase lottery ticket, d_2	$0	$0

a. A realistic estimate of the chances of winning is 1 in 250,000. Use the expected value approach to recommend a decision.
b. If a particular decision maker assigns an indifference probability of 0.000001 to the $0 payoff, would this individual purchase a lottery ticket? Use expected utility to justify your answer.

25. **Risk Profiles of Decision Makers.** Three decision makers have assessed utilities for the following decision problem (payoff in dollars):

DATA *file* riskprofile

Decision Alternative	State of Nature: s_1	s_2	s_3
d_1	20	50	−20
d_2	80	100	−100

The indifference probabilities are shown in the following table. **LO 8, 9**

Payoff	Indifference Probability (p): Decision Maker A	Decision Maker B	Decision Maker C
100	1.00	1.00	1.00
80	0.95	0.70	0.90

(*continued*)

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

50	0.90	0.60	0.75
20	0.70	0.45	0.60
−20	0.50	0.25	0.40
−100	0.00	0.00	0.00

a. Plot the utility function for money for each decision maker.
b. Classify each decision maker as a risk avoider, a risk taker, or risk-neutral.
c. For the payoff of 20, what is the premium that the risk avoider will pay to avoid risk? What is the premium that the risk taker will pay to have the opportunity of the high payoff?

26. **Recommended Decisions for Different Decision Makers.** In Problem 25, if $P(s_1) = 0.25$, $P(s_2) = 0.50$, and $P(s_3) = 0.25$, find a recommended decision for each of the three decision makers. *Note:* For the same decision problem, different utilities can lead to different decisions. **LO 3, 8**

27. **Calculating Utility for Exponential Utility Function.** Translate the following monetary payoffs into utilities for a decision maker whose utility function is described by an exponential function with $R = 250$: −$200, −$100, $0, $100, $200, $300, $400, $500. **LO 10**

28. **Decision Maker with an Exponential Utility Function.** Consider a decision maker who is comfortable with an investment decision that has a 50% chance of earning $25,000 and a 50% chance of losing $12,500, but not with any larger investments that have the same relative payoffs. **LO 9, 10**
    a. Write the equation for the exponential function that approximates this decision maker's utility function.
    b. Plot the exponential utility function for this decision maker for x values between −$20,000 and $35,000. Is this decision maker risk-seeking, risk-neutral, or risk-averse?
    c. Suppose the decision maker decides that they would actually be willing to make an investment that has a 50% chance of earning $30,000 and a 50% chance of losing $15,000. Plot the exponential function that approximates this utility function and compare it to the utility function from part (b). Is the decision maker becoming more risk-seeking or more risk-averse?

Case Problem 1: Property Purchase Strategy

Gina Foreman, president of Oceanview Development Corporation, is considering submitting a bid to purchase property that will be sold by sealed-bid auction at a county tax foreclosure. Gina's initial judgment is to submit a bid of $5 million. Based on Gina's experience, Gina estimates that a bid of $5 million will have a 0.2 probability of being the highest bid and securing the property for Oceanview. The current date is June 1. Sealed bids for the property must be submitted by August 15. The winning bid will be announced on September 1.

If Oceanview submits the highest bid and obtains the property, the firm plans to build and sell a complex of luxury condominiums. However, a complicating factor is that the property is currently zoned for single-family residences only. Gina believes that a referendum could be placed on the voting ballot in time for the November election. Passage of the referendum would change the zoning of the property and permit construction of the condominiums.

The sealed-bid procedure requires the bid to be submitted with a certified check for 10% of the amount bid. If the bid is rejected, the deposit is refunded. If the bid is accepted, the deposit is the down payment for the property. However, if the bid is accepted and the bidder does not follow through with the purchase and meet the remainder of the financial obligation within six months, the deposit will be forfeited. In this case, the county will offer the property to the next highest bidder.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

To determine whether Oceanview should submit the $5 million bid, Gina conducted some preliminary analysis. This preliminary work provided an assessment of 0.3 for the probability that the referendum for a zoning change will be approved and resulted in the following estimates of the costs and revenues that will be incurred if the condominiums are built:

Costs and Revenue Estimates

Revenue from condominium sales	$15,000,000
Costs	
Property	$5,000,000
Construction expenses	$8,000,000

If Oceanview obtains the property and the zoning change is rejected in November, Gina believes that the best option would be for the firm not to complete the purchase of the property. In this case, Oceanview would forfeit the 10% deposit that accompanied the bid.

Because the likelihood that the zoning referendum will be approved is such an important factor in the decision process, Gina suggested that the firm hire a market research service to conduct a survey of voters. The survey would provide a better estimate of the likelihood that the referendum for a zoning change would be approved. The market research firm that Oceanview Development has worked with in the past has agreed to do the study for $15,000. The results of the study will be available August 1, so that Oceanview will have this information before the August 15 bid deadline. The results of the survey will be a prediction either that the zoning change will be approved or that the zoning change will be rejected. After considering the record of the market research service in previous studies conducted for Oceanview, Gina developed the following probability estimates concerning the accuracy of the market research information:

$$P(A|s_1) = 0.9 \qquad P(N|s_1) = 0.1$$
$$P(A|s_2) = 0.2 \qquad P(N|s_2) = 0.8$$

where

A = prediction of zoning change approval
N = prediction that zoning change will not be approved
s_1 = the zoning change is approved by the voters
s_2 = the zoning change is rejected by the voters

Gina needs to consider all of this information in deciding how to move forward on behalf of Oceanview. **LO 1, 3, 6**

Managerial Report

Perform an analysis of the problem facing the Oceanview Development Corporation, and prepare a report that summarizes your findings and recommendations. Include the following items in your report:

1. A decision tree that shows the logical sequence of the decision problem
2. A recommendation regarding what Oceanview should do if the market research information is not available
3. A decision strategy that Oceanview should follow if the market research is conducted
4. A recommendation as to whether Oceanview should employ the market research firm, along with the value of the information provided by the market research firm

Include the details of your analysis as an appendix to your report.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Case Problem 2: Semiconductor Fabrication at Axeon Labs

Srinivas Pathan is the Chief Operating Officer for Axeon Labs, a manufacturer of integrated circuit (IC) chips and other components used in electronic devices. The IC chips produced by Axeon Labs are used to power computers, phones, and many other devices. The IC chips are produced through a semiconductor fabrication process. The process starts with blank discs made of silicon that are called wafers. The wafers undergo additional processing including etching and masking to create transistors that are extremely small in size (less than 10 nanometers) to make up the integrated circuit. Because of the exacting specifications of the processing steps, the IC chips are produced in clean rooms that control static electricity, particulate matter, and other potential sources of contamination. The clean rooms are subject to rigorous quality checks to maintain required quality. However, even with the exacting quality specifications, IC chips that are produced may not pass final inspection and cannot be sold to customers.

Srinivas has been approached by a new potential customer who is interested in purchasing IC chips from Axeon. This new customer would like to place a large one-time order. Srinivas must decide whether or not to accept the order and produce the IC chips for this new customer. Based on previous experience, Srinivas estimates that the probability the chips will be acceptable to the customer is 0.74, and the probability that the chips will be unacceptable to the customer is 0.26. For simplicity, we will define the event that the chips are acceptable as event A, and the event that the chips are unacceptable as event U, so $P(A) = 0.74$ and $P(U) = 0.26$. Srinivas estimates that if Axeon produces the chips and they are acceptable to the customer, this will generate a profit of \$2.3 million. But if Axeon produces the chips and they are unacceptable to the customer, then they will lose \$1.1 million. Alternatively, Srinivas can decide to not accept the order and not produce the IC chips, resulting in \$0 profit.

Srinivas also knows that before deciding whether or not to accept the order, Axeon can conduct a full quality check of its fabrication process. This quality check will give additional information on the likelihood of Axeon being able to produce acceptable or unacceptable chips for the customer. If Axeon chooses to do the quality check, they can either pass (event P) or fail (event F) the quality check. The revised probabilities for being able to produce IC chips that are acceptable/unacceptable are shown below.

$$P(A|P) = 0.93 \qquad P(A|F) = 0.11$$
$$P(U|P) = 0.07 \qquad P(U|F) = 0.89$$

If Axeon chooses to conduct the quality check, Srinivas believes that the extra cost will be \$15,000, and that the probability of failing the quality check is 0.3 ($P(F) = 0.3$). **LO 1, 3, 6**

Managerial Report

Perform an analysis of the decision facing Srinivas Pathan at Axeon Labs, and prepare a report that summarizes your findings and recommendations. Include the following items in your report:

1. A decision tree for this problem assuming Srinivas will first make a decision of whether or not to do the full quality check of its fabrication process and then will make the decision of whether to accept the order or not.
2. A recommendation of whether or not Srinivas should conduct the quality check of its fabrication process.
3. A recommendation for whether or not Srinivas should accept the order and produce the IC chips. If the recommendation is for Srinivas to first conduct the quality check of the fabrication process, then the recommendation for accepting the order will depend on the result of the quality check.

Include the details of your analysis as an appendix to your report.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Case Problem: Capital State University Game-Day Magazines

This case draws on material from Chapters 2, 3, 4, 8, 9, and 13.

Capital State University (CSU) is a leading Midwest University with a strong collegiate football program. Kris Stetzel serves as CSU's Associate Athletic Director for External Affairs. Kris' job responsibilities include negotiating with commercial vendors for services such as concessions at sporting events, event staff and security, and game-day hospitality. Kris brokers deals for corporate sponsorship of CSU athletic programs and arranges for radio and television coverage of CSU athletic events. Kris also manages CSU sports advertising and marketing and sports information-related media relations for print, radio, television, and online.

Recently, Kris has been examining CSU's business arrangement with the publishing company that prints the game-day sports magazines for CSU home football games. As part of a recent comprehensive university-wide sports media contract, CSU has a new publishing agreement with its print vendor. The magazines typically contain about 60 pages of information on the CSU football team and its opponent for that week. The magazines are sold at vendor stands positioned outside of CSU's football stadium.

Currently, CSU places one order in July, several months prior to the first home football game, that states how many magazines CSU wants for each home game of the season. The publishing company prints the magazines and ships all magazines to CSU prior to the first game of the season.

From data collected in past football seasons, Kris knows that CSU is often off by a considerable amount in their order quantities. Most weeks, CSU has many leftover magazines, but because the magazines are tailored to each home opponent, they cannot be resold in future weeks. In some weeks in previous football seasons, demand surpassed supply and CSU ran out of football magazines. Currently, CSU determines order quantities for each home game by looking at the past season's order quantities and then adjusting this amount up or down based on a gut feeling on how popular the current season's game would be in comparison to games in the previous season.

Kris believes that it should be possible to improve this ordering process. Kris has located data from the past nine football seasons. Kris has information on the following variables for each home game: the number of magazines sold, the year the game took place, the week that the game took place during the season, the opponent's preseason ranking, the number of preseason tickets sold for that game, the total game attendance, CSU's preseason rank, the number of the home game within CSU's season, whether or not the game was an in-conference game for CSU, whether or not the game was Homecoming for CSU, the game-day weather, the game-day kickoff temperature, the number of wins and losses for CSU's opponent in the previous season, and the number of wins and losses for CSU in the previous season. These data are in the file *magazines_csu*, and Figure Case.1 displays the data for Years 1 and 2. Kris also noted that the CSU game in Week 1 of Year 8 was somewhat of an anomaly because CSU wore special throwback uniforms to honor the players from their only National Championship season, which greatly increased attendance at that game.

DATA*file*
magazines_csu

Figure Case.1 Portion of Data Available for Use in Determining How Many CSU Football Magazines to Order

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Opponent	Magazine Sales (Units)	Year	Week In Season	Opponent Preseason Rank	Preseason Ticket Sales	Total Game Attendance	CSU Preseason Rank	Home Game Number	Conference Game (1 = Yes 0 = No)	Homecoming (1 = Yes 0 = No)	Game Day Weather	Kickoff Temperature	Opponent's Previous Season Number of Wins	Opponent's Previous Season Number of Losses	CSU's Previous Season Number of Wins	CSU's Previous Season Number of Losses
2	Cedar Falls University	4,165	1	2	120	47,420	66,515	21	1	0	0	Sunny	70	13	2	9	3
3	Oklahoma A&M	3,746	1	3	58	47,420	64,898	21	2	0	0	Rain	60	4	7	9	3
4	Urbana College	4,943	1	5	67	47,420	70,397	21	3	1	0	Sunny	68	2	9	9	3
5	University of Bloomington	2,366	1	9	83	47,420	69,185	21	4	1	1	Rain	42	3	8	9	3
6	Indiana A&M	1,796	1	10	74	47,420	70,397	21	5	1	0	Cloudy	46	5	8	9	3
7	Minneapolis State University	1,970	1	13	68	47,420	64,501	21	6	1	0	Cloudy	35	4	7	9	3
8	Mt Pleasant College	3,866	2	1	109	46,198	58,920	57	1	0	0	Sunny	82	2	9	7	5
9	University of Ames	5,194	2	2	99	46,198	70,397	57	2	0	0	Sunny	88	1	10	7	5
10	Ann Arbor University	1,900	2	5	6	46,198	70,397	57	3	1	0	Rain	52	12	0	7	5
11	Evanston University	2,523	2	6	55	46,198	70,397	57	4	1	1	Sunny	62	5	7	7	5
12	Madison University	2,734	2	8	17	46,198	70,397	57	5	1	0	Sunny	63	8	5	7	5
13	Columbus University	2,034	2	11	3	46,198	69,473	57	6	1	0	Sunny	61	10	3	7	5

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Kris has also done some investigation into the costs associated with ordering magazines from CSU's publisher. Under the current CSU contract with the publisher, CSU must determine order amounts for each upcoming home game by July. CSU pays $14.00 for each magazine that they order. Vendors then sell magazines at each CSU home football game for $25.00. CSU's agreement with the vendors states that CSU pays the vendor $2.50 for each magazine sold and keeps the remaining revenue. The current contract with the publisher states that the publisher must buy back any unsold magazines from CSU for $11.50.

Managerial Report

Use the concepts you have learned from Chapters 2, 3, 4, 8, 9, and 13 to write a report that will help Kris analyze football magazine sales in Years 1 through 9 to determine an order amount for Year 10. You should address each of the following in your report.

1. There is a considerable amount of data available in the file *magazines_csu*, but not all of it may be useful for your purposes here. Are there variables contained in the file *magazines_csu* that you would exclude from a forecast model to determine football magazine sales in Year 10? If so, why? Are there particular observations of football magazine sales from previous years that you would exclude from your forecasting model? If so, why?
2. Based on the data in the file *magazines_csu*, develop a regression model to forecast the average sales of football magazines for each of the seven home games in the upcoming season (Year 10). That is, you should construct a single regression model and use it to estimate the average demand for the seven home games in Year 10. In addition to the variables provided, you may create new variables based on these variables or based on observations of your analysis. Be sure to provide a thorough analysis of your final model (residual diagnostics) and provide assessments of its accuracy. What insights are available based on your regression model?
3. Use the forecasting model developed in part 2 to create a simulation model that Kris can use to estimate the total football magazine sales amounts in Year 10. Your simulation model should have seven uncertain inputs: one input for football magazine sales at each CSU game in Year 10. Then you should sum these sales amounts for each individual game to create a total football magazine sales amount for Year 10.
4. Kris has noticed that of the typical 60 pages in a football magazine, 45 of those 60 pages are the same for every game in a season. Only the 15 pages that discuss the weekly opponent change from week-to-week. CSU's publisher has indicated that it is possible for CSU to order generic game-day football magazines in the July preceding the season. This generic magazine contains the 45 pages of material that is the same for each game. Closer to the week of each game, CSU could then tailor the generic magazine with inserts specific to that week's game, along with a book jacket cover displaying players and coaches from the two teams playing that week. The number of game-specific inserts and book jacket covers can be determined closer to the actual games in order to allow for a more accurate forecast. Thus, the simulation model developed in part 3 effectively represents the sales amount for the generic magazine, and then CSU would order the game-specific inserts and book jacket covers much closer to the actual games when they have a much more accurate forecast of attendance and sales. However, Kris still is not sure how many generic magazines should be ordered. Should Kris order exactly the forecasted amount from part 3? More? Less? Why? Based on the cost values described from the publishing contract, if Kris orders 21,500 generic magazines in July, what are the estimated expected costs of lost sales (football magazines that CSU does not sell because they run out) and unsold magazines (football magazines that CSU must send back to the publisher at the end of the season)?
5. Assuming that CSU can tailor the specific magazines for each game in Year 10 at a later date, what is the optimal order amount for Kris to place in July prior to Year 10 for the generic magazines? The optimal order amount should minimize the total expected lost sales and unsold magazines cost in Year 10? Assume that Kris must order in batches of 500 magazines.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Hanover Inc.

This case draws on material from Chapters 12, 14, and 15.

Because of increasing global demand for its fiber optics products, Hanover Inc. is currently facing a decision on how to increase its production capacity. Hanover has three fiber optic products denoted by FA, FB, and FC. Hanover has two existing plants in Austin, Texas, and Paris, France. Hanover plans to keep the Austin and Paris plants open, but the production capacity at these two existing plants can be changed from their current levels of 600,000 units per year. Each plant can make any mix of the three fiber optic products. Seven locations are being considered for an additional new plant. These consist of two cities in the United States (Charleston, South Carolina and Mobile, Alabama) and five other locations currently simply defined by country (Australia, India, Malaysia, South Africa, and Spain).

The company's customers have been aggregated onto eight customer regions: Malaysia, China, France, Brazil, US Northeast, US Southeast, US Midwest, and US West. The Planning and Forecasting department has forecasted demand for each region five years into the future. The Planning and Forecasting department has also provided other relevant data to help with the decision on how to expand capacity given the forecasted demand. Descriptions of the available data provided in the file *hanover* follow:



- ***Duty:*** This worksheet contains the duty rate charged from each plant to each customer region. This rate is multiplied by the selling price to get the cost per unit paid in duty to ship into one country from another country.
- ***Fixed Cost:*** This worksheet contains the fixed cost of plant operation including capital costs, insurance, management, etc. These costs depend on the capacity level of the plant, and costs for two different levels of capacity are given for each plant location. The capacities are independent of product mix.
- ***Forecast:*** This worksheet contains five-year-out forecasted demand for each product by customer region.
- ***F&WH Cost:*** This worksheet contains freight and warehousing cost per unit from each plant location to each customer region (which are independent of product type).
- ***Price:*** This worksheet contains the selling price of each product for each customer region.
- ***Variable Cost:*** This worksheet contains the variable cost of each product at each plant.

The vice president of supply chain management at Hanover would like you to study this situation using the given data.

Managerial Report

Use the concepts in Chapters 12, 14, and 15 to make a recommendation to Hanover's vice president of supply chain management. In particular, make a recommendation regarding whether or not a new plant is needed and if so, at which of the seven locations under consideration it should be located. Your report should also indicate which capacity levels should be used at each plant, and how product should be sourced.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Appendix A—Basics of Excel

Contents

A.1 Using Microsoft Excel
- Basic Spreadsheet Workbook Operations
- Creating, Saving, and Opening Files in Excel

A.2 Spreadsheet Basics
- Cells, References, and Formulas in Excel
- Finding the Appropriate Excel Function
- Colon Notation
- Inserting a Function into a Worksheet Cell
- Using Relative versus Absolute Cell References

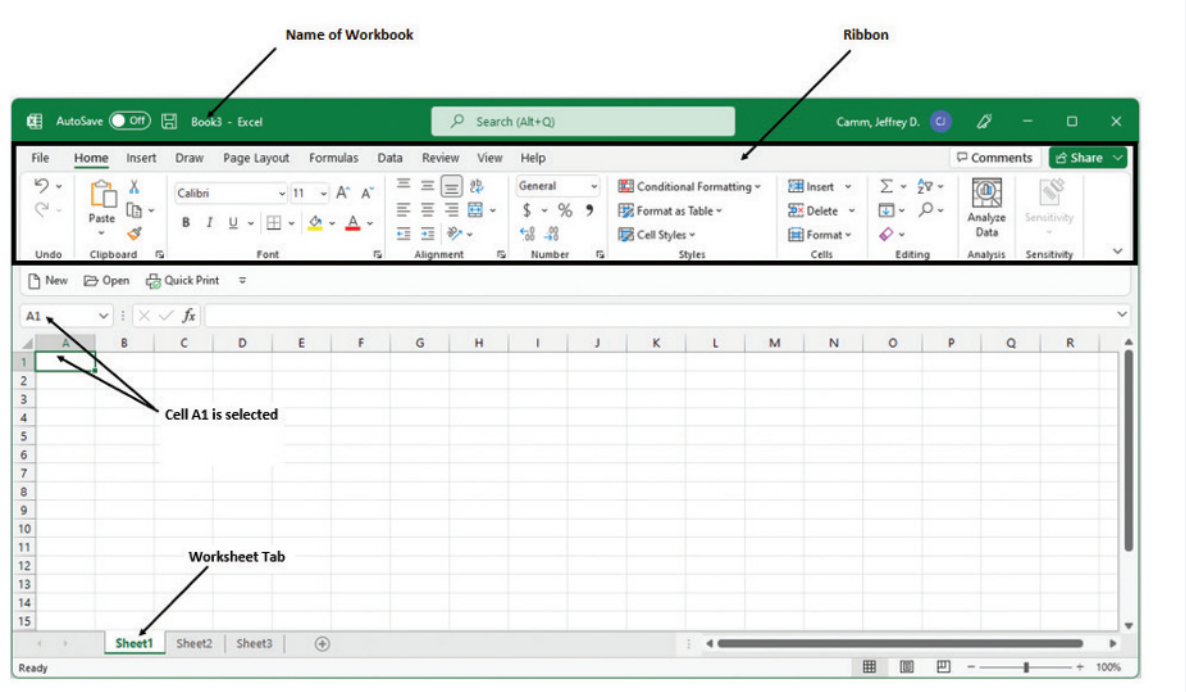
A.1 Using Microsoft Excel

Excel stores data and calculations in a file called a **workbook** that contains one or more **worksheets**. Figure A.1 shows the layout of a blank workbook created in Excel. The workbook is named Book3 and by default contains three worksheets named Sheet1, Sheet2 and Sheet3.

The wide bar located across the top of the workbook is referred to as the Ribbon. Tabs, located at the top of the Ribbon, contain groups of related commands. Ten tabs are included on the Ribbon in Excel: File, Home, Insert, Draw, Page Layout, Formulas, Data, Review, View and Help. Loading additional packages (such as Acrobat) may create additional tabs. Each tab contains several groups of related commands. The File tab is used to Open, Save, and Print files as well as to change the Options being used by Excel and to load Add-ins. Note that the Home tab is selected when a workbook is opened. Figure A.2 displays the ten groups located in the Home tab: Undo, Clipboard, Font, Alignment, Number, Styles, Cells, Editing, Analysis, and Sensitivty. Commands are arranged within each group.

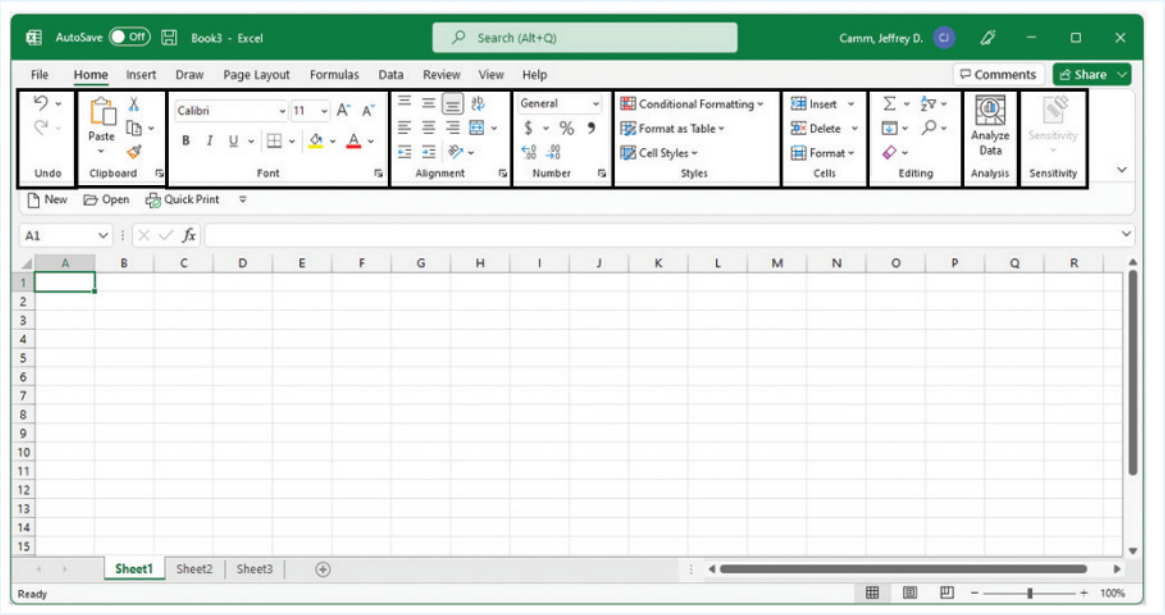
Depending on your installation, you may or may not see the Sensitivity group.

Figure A.1 Blank Workbook in Excel



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure A.2 Groups on the Home Tab in the Ribbon of an Excel Workbook

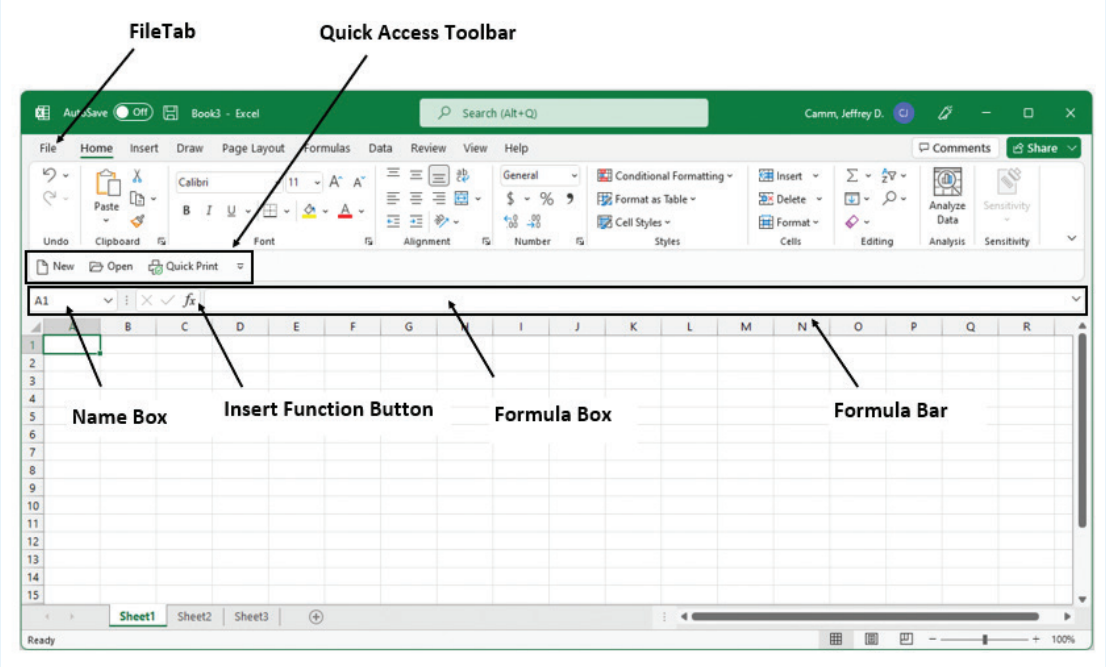


*Keyboard shortcut: pressing **Ctrl-B** will change the font of the text in the selected cell to bold. We include a full list of keyboard shortcuts for Excel at the end of this appendix.*

For example, to change selected text to boldface, click the **Home** tab and click the **Bold** button **B** in the **Font** group. The other tabs in the Ribbon are used to modify data in your spreadsheet or to perform analysis.

Figure A.3 illustrates the location of the File tab, the Quick Access toolbar, and the Formula bar. The Quick Access toolbar allows you to quickly access commonly used workbook functions.

Figure A.3 File Tab, Quick Access Toolbar, and Formula Bar of an Excel Workbook



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For instance, the Quick Access toolbar shown in Figure A.3 includes a **New** button New that can be used to open a new workbook without having to first click the **File** tab. To add or remove features on the Quick Access toolbar, click the **Customize Quick Access toolbar** button on the Quick Access toolbar.

The Formula bar contains a Name box, the Insert Function button *fx*, and a Formula box. In Figure A.3, "A1" appears in the Name box because cell A1 is selected. You can select any other cell in the worksheet by using the mouse to move the cursor to another cell and clicking or by typing the new cell location in the name box and pressing the Enter key. The Formula box is used to display the formula in the currently selected cell. For instance, if you had entered $=A1+A2$ into cell A3, whenever you select cell A3, the formula =A1+A2 will be shown in the Formula box. This feature makes it very easy to see and edit a formula in a cell. The Insert Function button allows you to quickly access all of the functions available in Excel. Later, we show how to find and use a particular function with the Insert Function button.

Basic Spreadsheet Workbook Operations

To change the name of the current worksheet, follow these steps:

Step 1. Right-click on the worksheet tab named **Sheet1**
Step 2. Select the **Rename** option
Step 3. Enter *Nowlin* to rename the worksheet and press **Enter**

You can create a copy of the newly renamed Nowlin worksheet by following these steps:

Step 1. Right-click the worksheet tab named **Nowlin**
Step 2. Select the **Move or Copy…** option
Step 3. When the **Move or Copy** dialog box appears, select the checkbox for **Create a Copy**, and click **OK**

The name of the copied worksheet will appear as "Nowlin (2)." You can then rename it, if desired, by following the steps outlined previously. Worksheets can also be moved to other workbooks or to a different position in the current workbook by using the Move or Copy option.

To create additional worksheets follow these steps:

New worksheets can also be created using the insert worksheet button ⊕ at the bottom of the screen.

Step 1. Right-click on the tab of any existing worksheet
Step 2. Select **Insert…**
Step 3. When the **Insert** dialog box appears, select **Worksheet** from the **General** area, and click **OK**

Worksheets can be deleted by right-clicking the worksheet tab and choosing **Delete**. After clicking **Delete**, a window may appear, warning you that any data appearing in the worksheet will be lost. Click **Delete** to confirm that you do want to delete the worksheet.

Creating, Saving, and Opening Files in Excel

To illustrate manually entering, saving, and opening a file, we will use the Nowlin Plastics make-versus-buy model from Chapter 12. The objective is to determine whether Nowlin should manufacture or outsource production for its Viper product next year. Nowlin must pay a fixed cost of $234,000 and a variable cost per unit of $2 to manufacture the product. Nowlin can outsource production for $3.50 per unit.

We begin by assuming that Excel is open and a blank worksheet is displayed. The Nowlin data can now be entered manually by simply typing the manufacturing fixed cost of $234,000, the variable cost of $2, and the outsourcing cost of $3.50 into the worksheet.

We will place the data for the Nowlin example in the top portion of Sheet1 of the new workbook. First, we enter the label *Nowlin Plastics* in cell A1 and click the **Bold** button in the **Font** group. Next, we enter the label *Parameters* and click on the **Bold** button in the **Font** group. To identify each of the three data values, we enter the label *Manufacturing*

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure A.4 Nowlin Plastics Data

	A	B	C
1	**Nowlin Plastics**		
2			
3	**Parameters**		
4	Manufacturing Fixed Cost	$234,000.00	
5	Manufacturing Variable Cost per Unit	$2.00	
6			
7	Outsourcing Cost per Unit	$3.50	
8			

Fixed Cost in cell A4, the label *Manufacturing Variable Cost per Unit* in cell A5, and the label *Outsourcing Cost per Unit* in cell A7. Next, we enter the actual data into the corresponding cells in column B: the value of *$234,000* in cell B4; the value of *$2* in cell B5; and the value of *$3.50* in cell B7. Figure A.4 shows a portion of the worksheet we have just developed.

Before we begin the development of the model portion of the worksheet, we recommend that you first save the current file; this will prevent you from having to reenter the data in case something happens that causes Excel to close. To save the workbook using the filename *nowlin*, we perform the following steps:

Step 1. Click the **File** tab on the Ribbon
Step 2. Click **Save** in the list of options
Step 3. Select **This PC** under **Save As**, and click **Browse**
Step 4. When the **Save As** dialog box appears:
Select the location where you want to save the file
Enter the filename *nowlin* in the **File name:** box
Click **Save**

Keyboard shortcut: To save the file, press ***Ctrl-S****.*

Excel's Save command is designed to save the file as an Excel workbook. As you work with and build models in Excel, you should follow the practice of periodically saving the file so that you will not lose any work. After you have saved your file for the first time, the Save command will overwrite the existing version of the file, and you will not have to perform Steps 3 and 4.

Sometimes you may want to create a copy of an existing file. For instance, suppose you change one or more of the data values and would like to save the modified file using the filename *nowlinmod*. The following steps show how to save the modified workbook using filename *nowlinmod*:

Step 1. Click the **File** tab in the Ribbon
Step 2. Click **Save As** in the list of options
Step 3. Select **This PC** under **Save As**, and click **Browse**
Step 4. When the **Save As** dialog box appears:
Select the location where you want to save the file
Type the filename *nowlinmod* in the **File name:** box
Click **Save**

Once the *nowlinmod* workbook has been saved, you can continue to work with the file to perform whatever type of analysis is appropriate. When you are finished working with the file, simply click the close-window button ✕ located at the top right-hand corner of the Ribbon.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Later, you can easily access a previously saved file. For example, the following steps show how to open the previously saved *nowlin* workbook:

Step 1. Click the **File** tab in the Ribbon
Step 2. Click **Open** in the list of options
Step 3. Select **This PC** under **Open** and click **Browse**
Step 4. When the **Open** dialog box appears:
Find the location where you previously saved the *nowlin* file
Click on the filename **nowlin** so that it appears in the **File name:** box
Click **Open**

A.2 Spreadsheet Basics

Cells, References, and Formulas in Excel

We begin by assuming that the *nowlin* workbook is open again and that we would like to develop a model that can be used to compute the manufacturing and outsourcing cost given a certain required volume. We develop the model based on the data in the worksheet shown in Figure A.4. The model will contain formulas that refer to the location of the data cells in the upper section of the worksheet. By putting the location of the data cells in the formula, we will build a model that can be easily updated with new data.

*To display all formulas in the cells of a worksheet, hold down the **Ctrl** key and then press the ~ key (usually located above the Tab key).*

To provide a visual reminder that the bottom portion of this worksheet will contain the model, we enter the label *Model* into cell A10 and press the **Bold** button in the **Font** group. In cell A11, we enter the label *Quantity*. Next, we enter the labels *Total Cost to Produce* in cell A13, *Total Cost to Outsource* in cell A15, and *Savings due to Outsourcing* in cell A17.

In cell B11 we enter *10,000* to represent the quantity produced or outsourced by Nowlin Plastics. We will now enter formulas in cells B13, B15, and B17 that use the quantity in cell B11 to compute the values for production cost, outsourcing cost, and savings from outsourcing. The total cost to produce is the sum of the manufacturing fixed cost (cell B4) and the manufacturing variable cost. The manufacturing variable cost is the product of the production volume (cell B11) and the variable cost per unit (cell B5). Thus, the formula for total variable cost is B11*B5; to compute the value of total cost, we enter the formula *=B4+B11*B5* in cell B13. Next, total cost to outsource is the product of the outsourcing cost per unit (cell B7) and the quantity (cell B11); this is computed by entering the formula *=B7*B11* in cell B15. Finally, the savings due to outsourcing is computed by subtracting the cost of outsourcing (cell B15) from the production cost (cell B13). Thus, in cell B17 we enter the formula *=B13-B15*. Figure A.5 shows the Excel worksheet values and formulas used for these calculations.

We can now compute the savings due to outsourcing by entering a value for the quantity to be manufactured or outsourced in cell B11. Figure A.5 shows the results after entering a value of 10,000 in cell B11. We see that a quantity of 10,000 units results in a production cost of $254,000 and outsourcing cost of $35,000. Thus, the savings due to outsourcing is $219,000.

Finding the Appropriate Excel Function

Excel provides a variety of built-in formulas or functions for developing mathematical models. If we know which function is needed and how to use it, we can simply enter the function into the appropriate worksheet cell. However, if we are not sure which functions are available to accomplish a task or are not sure how to use a particular function, Excel can provide assistance.

To identify the functions available in Excel, click the **Insert Function** button *fx* on the **Formula** bar; this opens the **Insert Function** dialog box shown in Figure A.6. The **Search for a function:** box at the top of the dialog box enables us to type a brief description of what we want to do. After entering a description and clicking **Go**, Excel will search for and display the functions that may accomplish our task in the **Select a function:** box. In many situations, however, we may want to browse through an entire category of functions to see

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

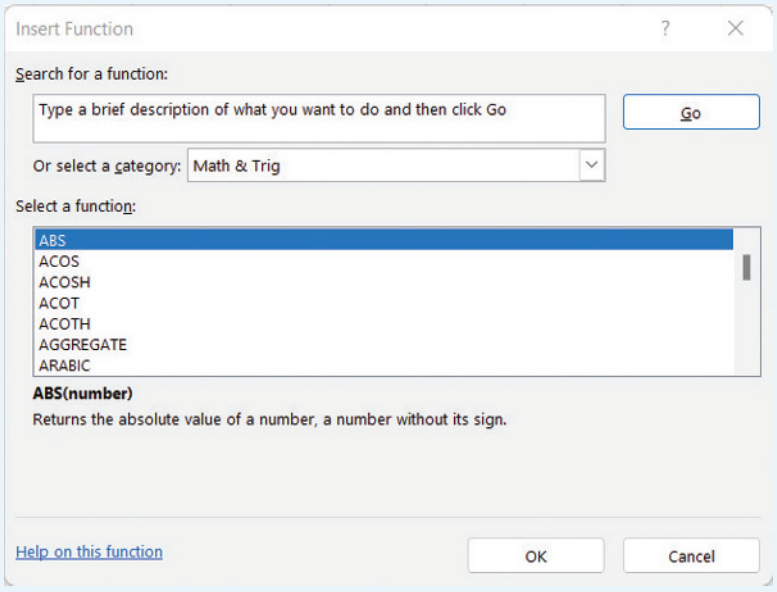
Figure A.5 Nowlin Plastics Data and Model

DATA *file* nowlin

	A	B	C
1	**Nowlin Plastics**		
2			
3	**Parameters**		
4	Manufacturing Fixed Cost	234000	
5	Manufacturing Variable Cost per Unit	2	
6			
7	Outsourcing Cost per Unit	3.5	
8			
9			
10	**Model**		
11	Quantity	10000	
12			
13	Total Cost to Produce	=B4+B11*B5	
14			
15	Total Cost to Outsource	=B7*B11	
16			
17	Savings due to Outsourcing	=B13-B15	
18			

	A	B	C
1	**Nowlin Plastics**		
2			
3	**Parameters**		
4	Manufacturing Fixed Cost	$234,000.00	
5	Manufacturing Variable Cost per Unit	$2.00	
6			
7	Outsourcing Cost per Unit	$3.50	
8			
9			
10	**Model**		
11	Quantity	10,000	
12			
13	Total Cost to Produce	$254,000.00	
14			
15	Total Cost to Outsource	$35,000.00	
16			
17	Savings due to Outsourcing	$219,000.00	
18			

Figure A.6 Insert Function Dialog Box



The ABS function calculates the absolute value of a number. The ACOS function calculates the arccosine of a number.

what is available. For this task, the **Or select a category:** box is helpful. It contains a drop-down list of several categories of functions provided by Excel. Figure A.6 shows that we selected the **Math & Trig** category. As a result, Excel's Math & Trig functions appear in alphabetical order in the **Select a function:** area. We see the ABS function listed first, followed by the ACOS function, and so on.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Colon Notation

Although many functions, such as the ABS function, have a single argument, some Excel functions depend on arrays. **Colon notation** provides an efficient way to convey arrays and matrices of cells to functions. For example, the colon notation B1:B5 means cell B1 "through" cell B5, namely the array of values stored in the locations (B1,B2,B3,B4,B5). Consider, for example, the following function =SUM(B1:B5). The sum function adds up the elements contained in the function's argument. Hence, =SUM(B1:B5) evaluates the following formula:

$$= B1 + B2 + B3 + B4 + B5$$

DATAfile nowlinfinancial

To illustrate the use of colon notation, we will consider the financial data for Nowlin Plastics contained in the DATAfile *nowlinfinancial* and shown in Figure A.7. Column A contains the name of each month, column B the revenue for each month, and column C the cost data. In row 15, we compute the total revenues and costs for the year. To do this we first enter *Total:* in cell A15. Next, we enter the formula *=SUM(B2:B13)* in cell B15 and *=SUM(C2:C13)* in cell C15. This shows that the total revenues for the company are $39,319,000 and the total costs are $36,549,000.

Inserting a Function into a Worksheet Cell

Continuing with the Nowlin financial data, we will now show how to use the Insert Function and Function Arguments dialog boxes to select a function, develop its arguments, and insert the function into a worksheet cell. We wish to calculate the average monthly revenue and cost at Nowlin. To do so, we execute the following steps:

*If you need additional guidance on the use of a particular function in Excel, the **Function Arguments** dialog box contains a link, **Help on this function**.*

Step 1. Select cell B17 in the DATAfile *nowlinfinancial*

Step 2. Click the **Insert Function** button *fx*.
Select **Statistical** in the **Or select a category:** box
Select **AVERAGE** from the **Select a function:** options

Figure A.7 Nowlin Plastics Monthly Revenues and Costs

	A	B	C
1	**Month**	**Revenue**	**Cost**
2	January	3459000	3250000
3	February	2873000	2640000
4	March	3195000	3021000
5	April	2925000	3015000
6	May	3682000	3150000
7	June	3436000	3240000
8	July	3410000	3185000
9	August	3782000	3237000
10	September	3548000	3196000
11	October	3136000	2997000
12	November	3028000	2815000
13	December	2845000	2803000
14			
15	**Total:**	=SUM(B2:B13)	=SUM(C2:C13)
16			
17	**Average:**	=AVERAGE(B2:B13)	=AVERAGE(C2:C13)

	A	B	C
1	**Month**	**Revenue**	**Cost**
2	January	$ 3,459,000	$ 3,250,000
3	February	$ 2,873,000	$ 2,640,000
4	March	$ 3,195,000	$ 3,021,000
5	April	$ 2,925,000	$ 3,015,000
6	May	$ 3,682,000	$ 3,150,000
7	June	$ 3,436,000	$ 3,240,000
8	July	$ 3,410,000	$ 3,185,000
9	August	$ 3,782,000	$ 3,237,000
10	September	$ 3,548,000	$ 3,196,000
11	October	$ 3,136,000	$ 2,997,000
12	November	$ 3,028,000	$ 2,815,000
13	December	$ 2,845,000	$ 2,803,000
14			
15	**Total:**	$ 39,319,000	$ 36,549,000
16			
17	**Average:**	$ 3,276,583	$ 3,045,750

DATAfile nowlinfinancial

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Step 3. When the **Function Arguments** dialog box appears:
Enter *B2:B13* in the **Number1** box
Click **OK**

Step 4. Repeat Steps 1 through 3 for the cost data in column C

Figure A.7 shows that the average monthly revenue is $3,276,583 and the average monthly cost is $3,045,750.

Using Relative versus Absolute Cell References

One of the most powerful abilities of spreadsheet software such as Excel is the ability to use relative references in formulas. Use of a **relative reference** allows the user to enter a formula once into Excel and then copy and paste that formula to other places so that the formula will update with the correct data without having to retype the formula. We will demonstrate the use of relative references in Excel by calculating the monthly profit at Nowlin Plastics using the following steps:

After completing Step 2, a shortcut to copying the formula to the range D3:D13 is to place the pointer in the bottom-right corner of cell D2 and then double-click.

*Keyboard shortcut: You can copy in Excel by pressing **Ctrl-C**. You can paste in Excel by pressing **Ctrl-V**.*

Step 1. Enter the label *Profit* in cell D1 and press the **Bold** button in the **Font** group of the **Home** tab

Step 2. Enter the formula =*B2-C2* in cell D2

Step 3. Copy the formula from cell D2 by selecting cell D2 and clicking **Copy** from the **Clipboard** group of the **Home** tab

Step 4. Select cells D3:D13

Step 5. Paste the formula from cell D2 by clicking **Paste** from the **Clipboard** group of the **Home** tab

The result of these steps is shown in Figure A.8, where we have calculated the profit for each month. Note that even though the only formula we entered was =*B2-C2* in cell D2, the formulas in cells D3 through D13 have been updated correctly to calculate the profit of each month using that month's revenue and cost.

Figure A.8 Nowlin Plastics Profit Calculation

	A	B	C	D
1	**Month**	**Revenue**	**Cost**	**Profit**
2	January	3459000	3250000	=B2-C2
3	February	2873000	2640000	=B3-C3
4	March	3195000	3021000	=B4-C4
5	April	2925000	3015000	=B5-C5
6	May	3682000	3150000	=B6-C6
7	June	3436000	3240000	=B7-C7
8	July	3410000	3185000	=B8-C8
9	August	3782000	3237000	=B9-C9
10	September	3548000	3196000	=B10-C10
11	October	3136000	2997000	=B11-C11
12	November	3028000	2815000	=B12-C12
13	December	2845000	2803000	=B13-C13
14				
15	**Total:**	=SUM(B2:B13)	=SUM(C2:C13)	
16				
17	**Average:**	=AVERAGE(B2:B13)	=AVERAGE(C2:C13)	

	A	B	C	D
1	**Month**	**Revenue**	**Cost**	**Profit**
2	January	$ 3,459,000	$ 3,250,000	$209,000
3	February	$ 2,873,000	$ 2,640,000	$233,000
4	March	$ 3,195,000	$ 3,021,000	$174,000
5	April	$ 2,925,000	$ 3,015,000	$ (90,000)
6	May	$ 3,682,000	$ 3,150,000	$532,000
7	June	$ 3,436,000	$ 3,240,000	$196,000
8	July	$ 3,410,000	$ 3,185,000	$225,000
9	August	$ 3,782,000	$ 3,237,000	$545,000
10	September	$ 3,548,000	$ 3,196,000	$352,000
11	October	$ 3,136,000	$ 2,997,000	$139,000
12	November	$ 3,028,000	$ 2,815,000	$213,000
13	December	$ 2,845,000	$ 2,803,000	$ 42,000
14				
15	**Total:**	$39,319,000	$36,549,000	
16				
17	**Average:**	$ 3,276,583	$ 3,045,750	

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In some cases, you may want Excel to use relative referencing for either the column or row location and absolute referencing for the other. For instance, to force Excel to always refer to column A but use relative referencing for the row, you would enter =$A1 into, say, cell B1. If this formula is copied into cell C3, the updated formula would be =$A3 (whereas it would be updated to =B3 if relative referencing was used for both the column and the row location).

In some situations, however, we do not want to use relative referencing in formulas. The alternative is to use an absolute reference, which we indicate to Excel by putting "$" before the row and/or column locations of the cell location. An **absolute reference** does *not* update to a new cell reference when the formula is copied to another location. We illustrate the use of an absolute reference by continuing to use the Nowlin financial data. Nowlin calculates an after-tax profit each month by multiplying its actual monthly profit by one minus its tax rate, which is currently estimated to be 30%. Cell B19 in Figure A.9 contains this tax rate. In column E, we calculate the after-tax profit for Nowlin in each month by using the following steps:

Step 1. Enter the label *After-Tax Profit* in cell E1 and press the **Bold** button in the **Font** group of the **Home** tab.
Step 2. Enter the formula *=D2*(1-$B$19)* in cell E2
Step 3. Copy the formula from cell E2 by selecting cell E2 and clicking **Copy** from the **Clipboard** group of the **Home** tab
Step 4. Select cells E3:E13
Step 5. Paste the formula from cell E2 by clicking **Paste** from the **Clipboard** group of the **Home** tab

Figure A.9 shows the after-tax profit in each month. Using $B$19 in the formula in cell E2 forces Excel to always refer to cell $B$19, even if we copy and paste this formula somewhere else in our worksheet. Notice that D2 continues to be a relative reference and is updated to D3, D4, and so on when we copy this formula to cells E3, E4, and so on, respectively.

Figure A.9 Nowlin Plastics After-Tax Profit Calculation Illustrating Relative versus Absolute References

	A	B	C	D	E
1	**Month**	**Revenue**	**Cost**	**Profit**	**After-Tax Profit**
2	January	3459000	3250000	=B2-C2	=D2*(1-$B$19)
3	February	2873000	2640000	=B3-C3	=D3*(1-$B$19)
4	March	3195000	3021000	=B4-C4	=D4*(1-$B$19)
5	April	2925000	3015000	=B5-C5	=D5*(1-$B$19)
6	May	3682000	3150000	=B6-C6	=D6*(1-$B$19)
7	June	3436000	3240000	=B7-C7	=D7*(1-$B$19)
8	July	3410000	3185000	=B8-C8	=D8*(1-$B$19)
9	August	3782000	3237000	=B9-C9	=D9*(1-$B$19)
10	September	3548000	3196000	=B10-C10	=D10*(1-$B$19)
11	October	3136000	2997000	=B11-C11	=D11*(1-$B$19)
12	November	3028000	2815000	=B12-C12	=D12*(1-$B$19)
13	December	2845000	2803000	=B13-C13	=D13*(1-$B$19)
14					
15	**Total:**	=SUM(B2:B13)	=SUM(C2:C13)		
16					
17	**Average:**	=AVERAGE(B2:B13)	=AVERAGE(C2:C13)		
18					
19	**Tax Rate:**	0.3			

	A	B	C	D	E
1	**Month**	**Revenue**	**Cost**	**Profit**	**After-Tax Profit**
2	January	$ 3,459,000	$ 3,250,000	$209,000	$ 146,300
3	February	$ 2,873,000	$ 2,640,000	$233,000	$ 163,100
4	March	$ 3,195,000	$ 3,021,000	$174,000	$ 121,800
5	April	$ 2,925,000	$ 3,015,000	$ (90,000)	$ (63,000)
6	May	$ 3,682,000	$ 3,150,000	$532,000	$ 372,400
7	June	$ 3,436,000	$ 3,240,000	$196,000	$ 137,200
8	July	$ 3,410,000	$ 3,185,000	$225,000	$ 157,500
9	August	$ 3,782,000	$ 3,237,000	$545,000	$ 381,500
10	September	$ 3,548,000	$ 3,196,000	$352,000	$ 246,400
11	October	$ 3,136,000	$ 2,997,000	$139,000	$ 97,300
12	November	$ 3,028,000	$ 2,815,000	$213,000	$ 149,100
13	December	$ 2,845,000	$ 2,803,000	$ 42,000	$ 29,400
14					
15	**Total:**	$39,319,000	$36,549,000		
16					
17	**Average:**	$ 3,276,583	$ 3,045,750		
18					
19	**Tax Rate:**	30%			

DATA*file*
nowlinfinancial_complete

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Summary

In Chapter 12, we give a detailed treatment of how to create more advanced business analytics models in Excel.

In this appendix, we have reviewed the basics of using Microsoft Excel. We have discussed the basic layout of Excel, file creation, saving, and editing as well as how to reference cells, use formulas, and use the copy and paste functions in an Excel worksheet. We have illustrated how to find and enter Excel functions and described the difference between relative and absolute cell references. We conclude this appendix with Table A.1, which shows commonly used keyboard shortcut keys in Excel. Keyboard shortcut keys can save considerable time when entering data into Excel.

Glossary

Absolute reference A reference to a cell location in an Excel worksheet formula that does not update according to its relative position when the formula is copied.
Colon notation Notation used in an Excel worksheet to denote "through." For example, =SUM(B1:B4) implies sum cells B1 through B4, or equivalently, B1 + B2 + B3 + B4.
Relative reference A reference to a cell location in an Excel worksheet formula that updates according to its relative position when the formula is copied.
Workbook An Excel file that contains a series of worksheets.
Worksheet A single page in an Excel workbook containing a matrix of cells defined by their column and row locations.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table A.1 Keyboard Shortcut Keys in Excel

Keyboard Shortcut Key	Task Description
Ctrl-S	Save
Ctrl-C	Copy
Ctrl-V	Paste
Ctrl-F	Find (can be used to find text both within a cell and within a formula in Excel)
Ctrl-P	Print
Ctrl-A	Selects all cells in the current data region
Ctrl-B	Changes the selected text to/from bold font
Ctrl-I	Changes the selected text to/from italic font
Ctrl-~ (usually located above the Tab key)	Toggles between displaying values and formulas in the Worksheet.
Ctrl-↓ (down arrow key)	Moves to the bottom-most cell of the current data region
Ctrl-↑ (up arrow key)	Moves to the top-most cell of the current data region
Ctrl-→ (right arrow key)	Moves to the right-most cell of the current data region
Ctrl-← (left arrow key)	Moves to the left-most cell of the current data region
Ctrl-Home	Moves to the top-left-most cell of the current data region
Ctrl-End	Moves to the bottom-left-most cell of the current data region
Shift-↓	Selects the current cell and the cell below
Shift-↑	Selects the current cell and the cell above
Shift-→	Selects the current cell and the cell to the right
Shift-←	Selects the current cell and the cell to the left
Ctrl-Shift-↓	Selects all cells from the current cell to the bottom-most cell of the data region
Ctrl-Shift-↑	Selects all cells from the current cell to the top-most cell of the data region
Ctrl-Shift-→	Selects all cells from the current cell to the right-most cell in the data region
Ctrl-Shift-←	Selects all cells from the current cell to the left-most cell in the data region
Ctrl-Shift-Home	Selects all cells from the current cell to the top-left-most cell in the data region
Ctrl-Shift-End	Selects all cells from the current cell to the bottom-right-most cell in the data region
Ctrl-Spacebar	Selects the entire current column
Shift-Spacebar	Selects the entire current row

A data region refers to all adjacent cells that contain data in an Excel worksheet.

Holding down the **Ctrl** *key and clicking on multiple cells allows you to select multiple nonadjacent cells. Holding down the* **Shift** *key and clicking on two nonadjacent cells selects all cells between the two cells.*

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Appendix B—Database Basics with Microsoft Access

Contents

B.1 Database Basics
- Considerations When Designing a Database
- Creating a Database in Access

B.2 Creating Relationships Between Tables in Microsoft Access

B.3 Sorting and Filtering Records

B.4 Queries
- Select Queries
- Action Queries
- Crosstab Queries

B.5 Saving Data to External Files

Data are the cornerstone of analytics; without accurate and timely data on relevant aspects of a business or organization, analytic techniques are useless, and the resulting analyses are meaningless (or worse yet, potentially misleading). The data used by organizations to make decisions are not static, but rather are dynamic and constantly changing, usually at a rapid pace. Every change or addition to a database represents a new opportunity to introduce errors into the data, so it is important to be capable of searching for duplicate entries or entries with errors. Furthermore, related data may be stored in different locations to simplify data entry or increase security. Because an analysis frequently requires information from several sets of data, an analyst must be able to efficiently combine information from multiple data sets in a logical manner. In this appendix, we will review tools in Microsoft Access® that can be used for these purposes.

B.1 Database Basics

A **database** is a collection of logically related data that can be retrieved, manipulated, and updated to meet a user's or organization's needs. By providing centralized access to data efficiently and consistently, a database serves as an electronic warehouse of information on some specific aspect of an organization. A database allows for the systematic accumulation, management, storage, retrieval, and analysis of the information it contains while reducing inaccuracies that routinely result from manual record keeping. Organizations of all sizes maintain databases that contain information about their customers, markets, suppliers, and employees. Before embarking on designing a database, it is important to consider what are good characteristics of a database. Foremost, the information in a database should be correct and complete so that decisions based on reports retrieved from the database will be based on accurate information. Second, a database should avoid duplicate information as much as possible in order to minimize wasted space and reduce the likelihood of errors and inconsistencies. Thus, a good database design

- divides the organization's information into subject-based tables to reduce redundant data without loss of information.
- provides the organization's database software with the information required to join information in tables together as needed.
- supports, maintains, and ensures the integrity and accuracy of the organization's information.
- avoids tables that have large numbers of entries with empty attributes.
- protects the organization's information through database security.
- accommodates the organization's data processing and reporting needs.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

We introduce Stinson's MicroBrew Distributor in Chapter 4.

Throughout this appendix, we will consider issues that arise in the creation and maintenance of a database for Stinson's MicroBrew Distributor, a licensed regional independent distributor of beer and a member of the National Beer Wholesalers Association. Stinson's provides refrigerated storage, transportation, and delivery of premium beers produced by several local microbreweries, so the company's facilities include a state-of-the-art temperature-controlled warehouse and a fleet of temperature-controlled trucks. Stinson's also employs sales, receiving, warehousing/inventory, and delivery personnel. When making a delivery, Stinson's monitors the retailer's shelves, taps, and keg lines to ensure the freshness and quality of the product. Because beer is perishable and because microbreweries often do not have the capacity to store, transport, and deliver large quantities of the products they produce, Stinson's holds a critical position in this supply chain.

Stinson's needs to develop a faster, more efficient, and more accurate means of recording, maintaining, and retrieving data related to various aspects of its business. The company's management team has identified three broad key areas of data management: personnel (information on Stinson's employees); supplier (information on purchases of beer made by Stinson's from its suppliers); and retailer (information on sales to Stinson's retail customers). We will use Microsoft Access 365 in designing Stinson's database. Access is a *relational* database management system (RDBMS), which is the most commonly used type of database system in business. Data in a relational database are stored in tables, which are the fundamental components of a database. A relational database allows the user to retrieve subsets of data from tables and retrieve and combine data that are stored in related tables.

In this section we will learn how to use Access to create a database and perform some basic database operations. In Access, a database is defined as a collection of related objects that are saved as a single file. An object in Access can be a:

- **Table**: Data arrayed in rows and columns (similar to a worksheet in an Excel spreadsheet) in which rows correspond to **records** (the individual units from which the data have been collected) and columns correspond to **fields** (the variables on which data have been collected from the records)
- **Form**: An object that is created from a table to simplify the process of entering data
- **Query**: A question posed by a user about the data in the database
- **Report**: Output from a table or a query that has been put into a specific prespecified format

We will focus on tables and queries in this appendix. You can refer to a wide variety of books on database design to learn about forms, reports, and other database objects.

Tables are the foundation of an Access database. Each field in a table has a data type. The most commonly used are as follows:

- *Short Text:* A field that contains words (such as the field *EmpFirstName* that is used to record the first name of the Stinson's employee); can contain no more than 255 alphanumeric characters.
- *Long Text:* A larger field that contains words and is generally used for recording lengthy descriptive entries (such as the field *Notes on Special Circumstances for a Transaction* that may be used to record detailed notes about unique aspects of specific transactions between Stinson's and its retail customers); can contain up to approximately 1 gigabyte, but controls to display a long text are limited to the first 64,000 characters.
- *Number*: A field that contains numerical values. There are several sizes of Number fields, which include:
  - *Byte:* Stores whole numbers from 0 to 255
  - *Decimal:* Stores numbers from $-10^{28} + 1$ to $10^{28} - 1$
  - *Integer:* Stores nonfractional numbers from $-32{,}768$ to $32{,}767$
  - *Long Integer:* Stores nonfractional numbers from $-2{,}147{,}483{,}648$ to $2{,}147{,}483{,}647$

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- *Currency:* A field that contains monetary values (such as the field *Transaction Amount* that may be used to record payments for goods that have been ordered by Stinson's retail customers).
- *Yes/No:* A field that contains binary variables (such as the field *Sunday Deliveries?* that may be used to record whether Stinson's retail customers accept deliveries on Sundays).
- *Date/Time:* A field that contains dates and times (such as the field *Date of Order* that may be used to record the date of an order placed by Stinson's with one of its suppliers).

Once you create a field and set its data type, you can set additional field properties. For example, for a numerical field you can define the data size to be Byte, Integer, Long Integer, Single, Double, Replication ID, or Decimal.

A Replication ID field is used for storing a globally unique identifier to prevent duplication of an identifier (such as customer number) when multiple copies of the same database are in use in different locations.

A database may consist of several tables that are maintained separately for a variety of reasons. We have already mentioned that Stinson's maintains information on its personnel, its suppliers and orders and its retail customers and sales. With regard to its retail customers, Stinson's may maintain information on the company name, street address, city, state, zip code, telephone number, and e-mail address; the dates of orders placed and quantities ordered; and the dates of actual deliveries and quantities delivered. In this example, we may consider establishing a table on Stinson's retailer customers; in this table each record corresponds to a retail customer, and the fields include the retail customer's company name, street address, city, state, zip code, telephone number, and e-mail address. Maintenance of this table is relatively simple; these data likely are not updated frequently for existing retail customers, and when Stinson's begins selling to a new retail customer, it has to establish only a single new record containing the information for the new retail customer in each field.

Stinson's may maintain other tables in this database. To track purchases made by its retail customers, the company may maintain a table of retail orders that includes the retail customer's name and the dollar value, date, and number of kegs and cases of beer for each order received by Stinson's. Because this table contains one record for each order placed with Stinson's, this table must be updated much more frequently than the table of information on Stinson's retailer customers.

In addition to answering a user's questions about the data in one or more tables, a query can also be used to add a record to the end of a table, delete a record from a table, or change the values for one or more records in a table. These functions are accomplished through append, delete, and update queries. We discuss queries in more detail later in this appendix.

A user who submits a query is effectively asking a question about the information in one or more tables in a database. For example, suppose Stinson's has determined that it has surplus kegs of Fine Pembrook Ale in inventory and is concerned about potential spoilage. As a result, the Marketing Department decides to identify all retail customers who have ordered kegs of Fine Pembrook Ale during the previous three months so that Stinson's can call these retailers and offer them a discounted price on additional kegs of this beer. A query could be designed to search the Retail Orders table for retail customers who meet this criterion. When the query is run, the output of the query provides the answer.

More complex queries may require data to be retrieved from multiple tables. For these queries, the tables must be connected by a join operation that links the records of the tables by their values in some common field. The common field serves as a bridge between the two tables, and the bridged tables are then treated by the query as a large single table comprising the fields of the original tables that have been joined. In designing a database for Stinson's, we may include the customer ID as a field in both the table of retail customers and the table of retail orders; values in the field customer ID would then provide the basis for linking records in these two tables. Thus, even though the table of retail orders does not contain the information on each of Stinson's retail customers that is contained in the table of Stinson's retail customers, if the database is well designed, the information in these two tables can easily be combined whenever necessary.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Each table in a database generally contains a **primary key field** that has a unique value for each record in the table. A primary key field is used to identify how records from several tables in a database are logically related. In our previous example, Customer ID is the primary key field for the table of Stinson's retail customers. To facilitate the linking of records in the table of Stinson's retail customers with logically related records in the table of retail orders, the two tables must share a primary key. Thus, a field for Customer ID may be included in the table of retail orders so that information in this table can be linked to information on each of Stinson's retail customers; when a field is included in a table for the sole purpose of facilitating links with records from another table, the field is referred to as a **foreign key field**. For tables that do not include a primary key field, a unique identifier for each record in the table may be formed by combining two or more fields. If the combination of these two fields will yield a unique value for each record that may be included in the table, the result is called a **compound primary key** and it is used in the same way a primary key is used.

Considerations When Designing a Database

Before creating a new database, we should carefully consider the following issues:

- What is the purpose of this database?
- Who will use this database?
- What queries and reports do the users of this database need?
- What information or data (fields) will this database include?
- What tables must be created, and how will the fields be allocated to these tables?
- What are the relationships between these tables?
- What are the fields that will be used to link related tables?
- What forms does the organization need to create to support the use of this database?

The answers to these questions will enable us to efficiently create a more effective and useful database. Let us consider these issues within the context of designing Stinson's database. Stinson's has several reasons for developing and implementing a database. Quick access to reliable and current data will enable Stinson's to monitor inventory and place orders from the microbreweries so that it can meet the demand of the retailers it supplies, while avoiding excess quantities and potential spoilage of inventory. These data can also be used to monitor the age of the product in inventory, which is a critical issue for a perishable product. Patterns in the orders of various beers placed by Stinson's retail customers can be analyzed to determine forecasts of future demand. Employees' salaries, federal and state tax withholding, vacation and sick days taken/remaining for the current year, and contributions to retirement funds can be tracked. Orders received from retail customers and Stinson's deliveries can be better coordinated. In summary, Stinson's can use a database to utilize information about its business in numerous ways that will potentially improve the efficiency and profitability of the company.

If we were to create a database for Stinson's MicroBrew Distributor, who within the company might need to use information from the database? A quick review of Stinson's reasons for developing and implementing a database provides the answer. Warehousing/inventory can use the database to control inventory. Delivery can create efficient delivery routes for the drivers on a daily basis and assess the on-time performance of the delivery system. Receiving can anticipate and prepare to receive daily deliveries of microbrews. Human resources can administer payroll, taxes, and benefits. Marketing can identify and exploit potential sales opportunities.

By considering the users and uses for the database, we can make a preliminary determination of the queries and reports the users of this database will need and the data (fields) this database must include. At this point we can consider the tables to be created, how the fields will be allocated to the tables, and the potential relationships between these tables. We can see that we will need to incorporate data on:

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- each microbrewery for which Stinson's distributes beer (Stinson's suppliers).
- each order placed with and delivery received from the microbreweries (Stinson's supplies).
- each retailer to which Stinson's distributes beer (Stinson's customers).
- each order received from and delivery made to Stinson's retail customers (Stinson's sales).
- each of Stinson's employees (Stinson's workforce).

As we design these tables and allocate fields to the tables we design, we must ensure that our database stores Stinson's data in the correct formats and is capable of outputting the queries, forms, and reports that Stinson's employees need.

With these considerations in mind, we decide to begin with the following 11 tables and associated fields in designing a database for Stinson's MicroBrew Distributor:

- TblEmployees
  - EmployeeID
  - EmpFirstName
  - EmpLastName
  - DOB
  - Street Address
  - City
  - State
  - Zip Code
  - Phone Number
- TblJobTitle
  - Job ID
  - Job Title
- TblEmployHist
  - EmployeeID
  - Start Date
  - End Date
  - Job ID
  - Salary
  - Hourly Rate
- TblBrewers
  - BrewerID
  - Brewery Name
  - Street Address
  - City
  - State
  - Zip Code
  - Phone Number
- TblSOrders
  - SOrder Number
  - BrewerID
  - Date of SOrder
  - EmployeeID
  - Keg or Case?
  - SQuantity Ordered
- TblSDeliveries
  - SOrder Number
  - BrewerID
  - EmployeeID
  - Date of SDelivery
  - SQuantity Delivered
- TblPurchasePrices
  - BrewerID
  - KegPurchasePrice
  - CasePurchasePrice
- TblRetailers
  - CustID
  - Name
  - Class
  - Street Address
  - City
  - State
  - Zip Code
  - Phone Number

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

- TblROrders
  - ROrder Number
  - Name
  - CustID
  - BrewerID
  - Date of ROrder
  - Keg or Case?
  - RQuantity Ordered
  - Rush?
- TblRDeliveries
  - CustID
  - Name
  - ROrder Number
  - EmployeeID
  - Date of RDelivery
  - RQuantity Delivered
- TblSalesPrices
  - BrewerID
  - KegSalesPrice
  - CaseSalesPrice

Each table contains information about a particular aspect of Stinson's business operations:

- *TblEmployees*: Information about each Stinson's employee, primarily obtained when the employee is hired
- *TblJobTitle*: Information about each type of position held by Stinson's employees
- *TblEmployHist*: Information about the employment history of each Stinson's employee
- *TblBrewers*: Information about each microbrewery that supplies Stinson's with beer
- *TblSOrders*: Information about each order that Stinson's has placed with the microbreweries that supply Stinson's with beer
- *TblSDeliveries*: Information about each delivery that Stinson's has received from the microbreweries that supply Stinson's with beer
- *TblPurchasePrices*: Information about the price charged by each microbrewery that supplies Stinson's with beer
- *TblRetailers*: Information about each retailer that Stinson's supplies with beer
- *TblROrders*: Information about each order that Stinson's has received from the retailers that Stinson's supplies with beer
- *TblRDeliveries*: Information about each delivery that Stinson's has made to the retailers that Stinson's supplies with beer
- *TblSalesPrices*: Information about the price charged to retailers by Stinson's for each of the microwbrews it distributes

Note that the name of each table begins with the three letter designation Tbl*; this is consistent with the Leszynski/Reddick guidelines, a common set of standards for naming database objects.*

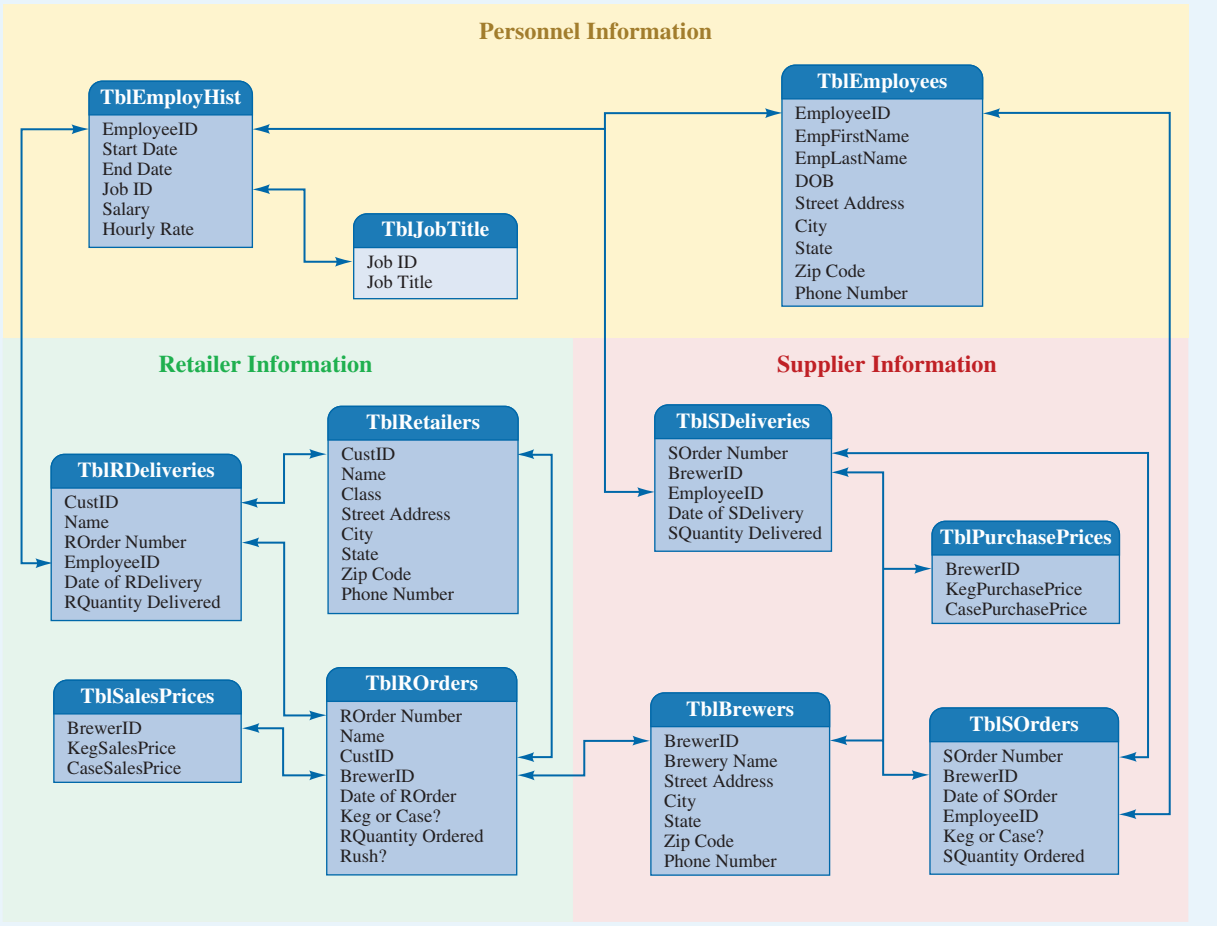
The first three tables deal with personnel information, the next four with product supply/purchasing information, and the last four with demand/sales information. Figure B.1 shows how these tables are related.

The relationships among the tables define how they can be linked. For example, suppose Stinson's Shipping Manager needs information on the orders placed by Stinson's retail customers that are to be filled tomorrow so that she can solve an optimization model that provides the optimal routes for Stinson's delivery trucks. The Shipping Manager needs to generate a report that includes the amount of various beers ordered and the address of each retail customer that has placed an order. To do so, she can use the common field CustID to link records from the TblRetailers. When the delivery has been made, the relevant information is input into the TblRDeliveries table. If the Shipping Manager needs to generate a report of deliveries made by each driver for the past week, she can use the common field EmployeeID to link records from the TblEmployees table with related records from the TblRDeliveries table.

Once Stinson's is satisfied that the planned database will provide the organization with the capability to collect and manage its data, and Stinson's is also confident that the database is capable of outputting the queries, forms, and reports that its employees need, we can proceed by using Access to create the new database. However, it is important to

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.1 Tables and Relationships for Stinson's Microbrew Distributor Database



realize that it is unusual for a new database to meet all of the potential needs of its users. A well-designed database allows for augmentation and revision when needs that the current database does not meet are identified.

Creating a Database in Access

When you open Access, the left pane provides links to databases you have recently opened as well as a means for opening existing database documents. The available document templates are provided in the right pane; these preinstalled templates can be used to create new databases that utilize common formats. Because we are focusing on building a fairly generic database, we will use the Blank desktop database tool. We are now ready to create a new database by following these steps:

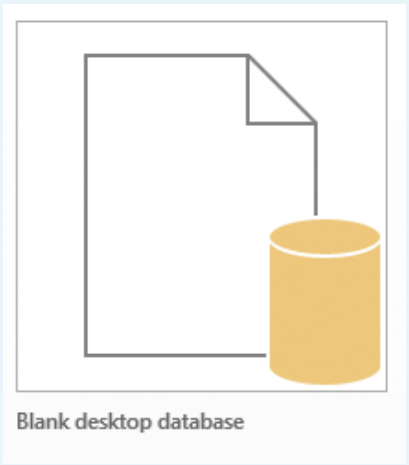
Step 1. Click the **Blank desktop database** icon (Figure B.2)

Step 2. When the **Blank desktop database** dialog box (Figure B.3) appears:
Enter the name of the new database in the **File Name** box (we will call our new database *stinsons.accdb*)
Indicate the location where the new database will be saved by clicking the **Browse** button (we will save *stinsons.accdb* in the folder *C:\Stinson Files*)

Step 3. Click **Create**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.2 Blank Desktop Database Icon

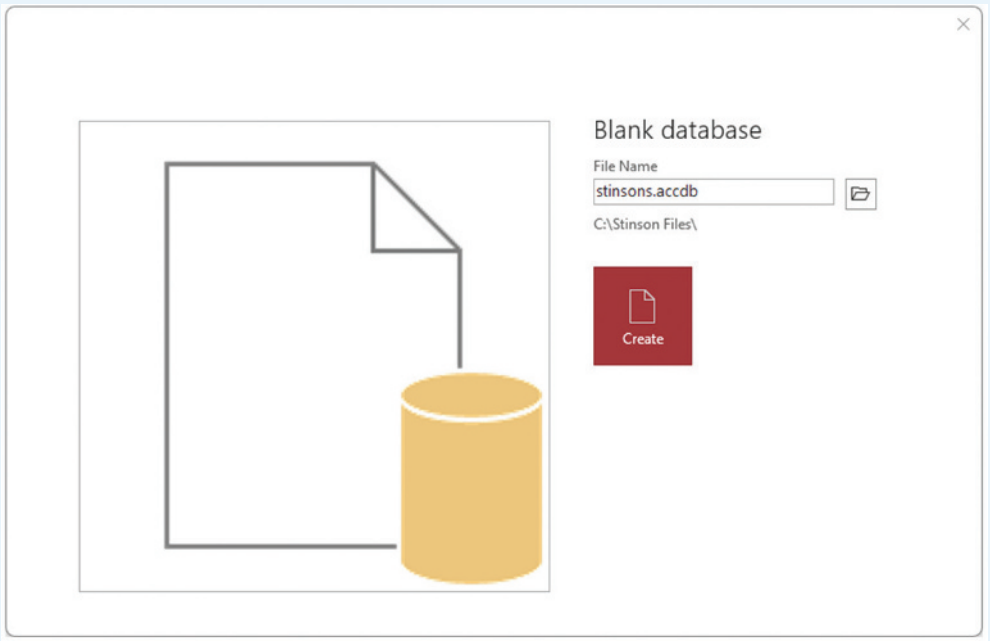


*Clicking the **File** tab in the Ribbon will allow the user to create new databases and access existing databases from the Datasheet view.*

This takes us to the Access Datasheet view. As shown in Figure B.4, the Datasheet view includes a Navigation Panel and Table Window. The Ribbon in the Datasheet view contains the **File, Home, Create, External Data, Database Tools, Fields**, and **Table** tabs.

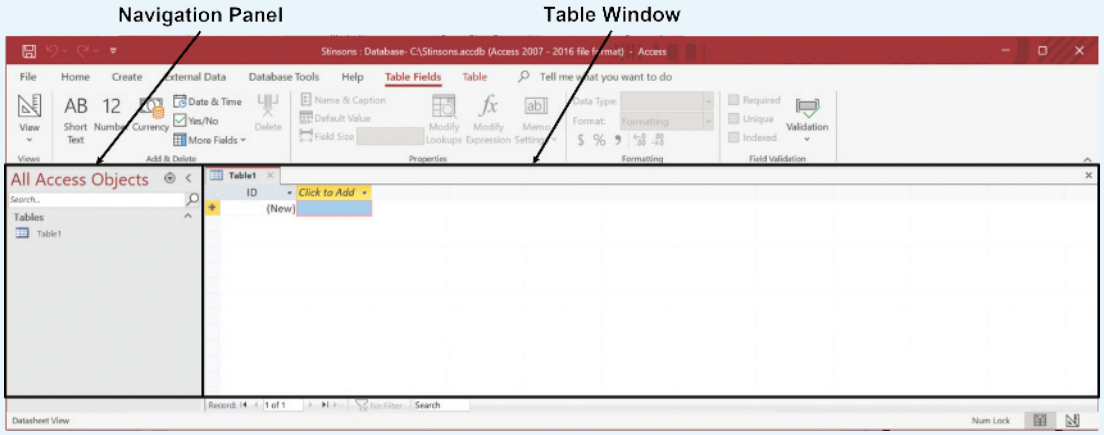
The **Datasheet view** provides the means for controlling the database. The groups and buttons of the Table Tools contextual tab are displayed across the top of this window. The Navigation Panel on the left side of the display lists all objects in the database. This

Figure B.3 Blank Desktop Database Dialog Box



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.4 Datasheet View and Table Tools Contextual Tab



provides a user with direct access to tables, reports, queries, forms, and so on that make up the currently open database. On the right side of the display is the Table Window; the tab in the upper left corner of the Table Window shows the name of the current table (Table1 in Figure B.4). In the Table Window, we can enter data directly into the table or modify data in an existing table.

The first step in creating a new database is to create one or more tables. Because tables store the information contained in a database, they are the foundation of a database and must be created prior to the creation of any other objects in the database. There are two options for manually creating a table: We can enter data directly in Datasheet view, or we can design a table in **Design view**. We will create our first table, TblBrewers, by entering data directly in Datasheet view. You can review an example database comprising all of the objects and relationships between the objects that we create throughout this appendix for the Stinson's database in the file *stinsons.accdb*.

MODEL *file*
stinsons.accdb

In Datasheet view the data are entered by field, one record at a time. In Figure B.1 we see that the fields for TblBrewers are BrewerID, Brewery Name, Street Address, City, State, Zip Code, and Phone Number. From Stinson's current filing system, we have been able to retrieve the information in Table B.1 on the breweries that supply Stinson's.

We can enter these data into our new database in Datasheet view by following these steps:

When we enter 3 in Step 1, this establishes a new field with the generic name "Field1" and generates a value for the ID column. Pressing the ***Tab*** *key moves to the next field entry box for the same record.*

Step 1. Enter the first record from Table B.1 into the first row of the **Table Window** in Access by entering a *3* in the top row next to **(New)**, pressing the **Tab** key, entering *Oak Creek Brewery* in the next column, pressing the **Tab** key, entering *12 Appleton St* in the next column, pressing the **Tab** key, entering *Dayton* in the next column, pressing the **Tab** key, and so on.

Step 2. Enter the second record from Table B.1 by repeating Step 1 for the Gonzo Microbrew data and entering these data into the second row of the **Table Window** in Access

Continue entering data for the remaining microbreweries in this manner

The completed table in Access appears in Figure B.5.

Now that we have entered all of our information on the microbreweries that supply Stinson's, we need to save this table as an object in the Stinson's database. We click

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table B.1 Raw Data for Table TblBrewers

BrewerID	Brewery Name	Street Address	City	State	Zip Code	Phone Number
3	Oak Creek Brewery	12 Appleton St	Dayton	OH	45455	937-449-1212
6	Gonzo Microbrew	1515 Marlboro Lane	Dayton	OH	45429	937-288-2651
4	McBride's Pride	425 Dry River Rd	Miamisburg	OH	45459	937-441-0123
9	Fine Pembrook Ale	141 Dusselberg Ave	Trotwood	OH	45426	937-852-8752
7	Midwest Fiddler Crab	844 Small Hills Ave	Kettering	OH	45453	937-633-7183
2	Herman's Killer Brew	912 Highlite Dr	Fairborn	OH	45442	937-883-2651

Figure B.5 Records for Six Microbreweries Entered into an Access Table

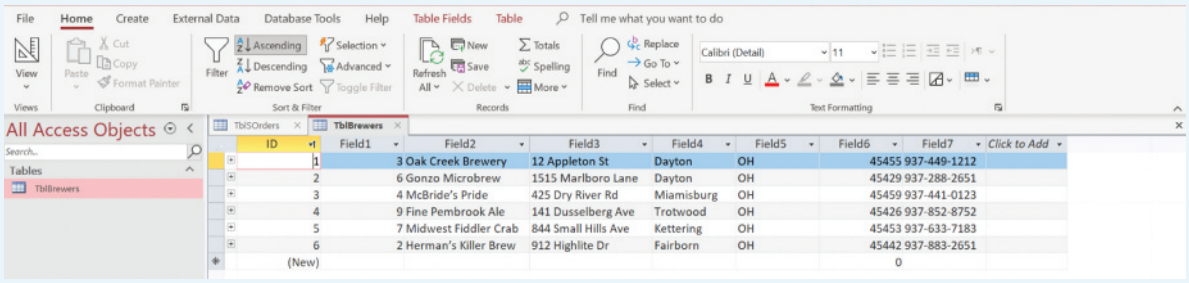
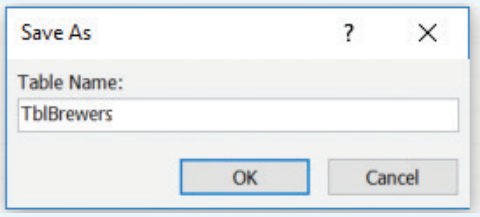


Figure B.6 Save as Dialog Box



the **Save** button in the **Quick Access Toolbar** above the Ribbon, type the table name *TblBrewers* in the **Save As** dialog box that appears (as shown in Figure B.6), and click **OK**. The name in the Table Name tab on the Table Window now reads "TblBrewers."

We can now use the Design view to provide meaningful names for our fields and specify each field's properties. We switch to Design view by first clicking on the arrow below the **View** button in the **Views** group of the Ribbon. This will open a pull-down menu with options for various views (recall that we are currently in the Datasheet view). Clicking on

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the **Design View** option opens the Design view for the current table as shown in Figure B.7. From the Design view we can define or edit the table's fields and field properties as well as rearrange the order of the fields if we wish. The name of the current table is again identified in the Name tab, and the Table Window is replaced with two sections: the Table Design Grid on top and the Field Properties Pane on the bottom of this window.

We can now replace the generic field names (Field1, Field2, etc.) in the column on the right side of the Table Design Grid of TblBrewers with the names we established from our original database design and then move to defining the field type for each field. To change the data type for a field in design view, we follow these steps:

Note that Field Names used in Access cannot exceed 64 characters, cannot begin with a space, and can include any combination of letters, numbers, spaces, and special characters except for a period (.), an accent grave (`), an exclamation point (!), or square brackets ([and]).

Step 1. Click the cell in the **Data Type** column (the middle column) in the **Table Design Grid** in the row of the field for which you want to change the data type

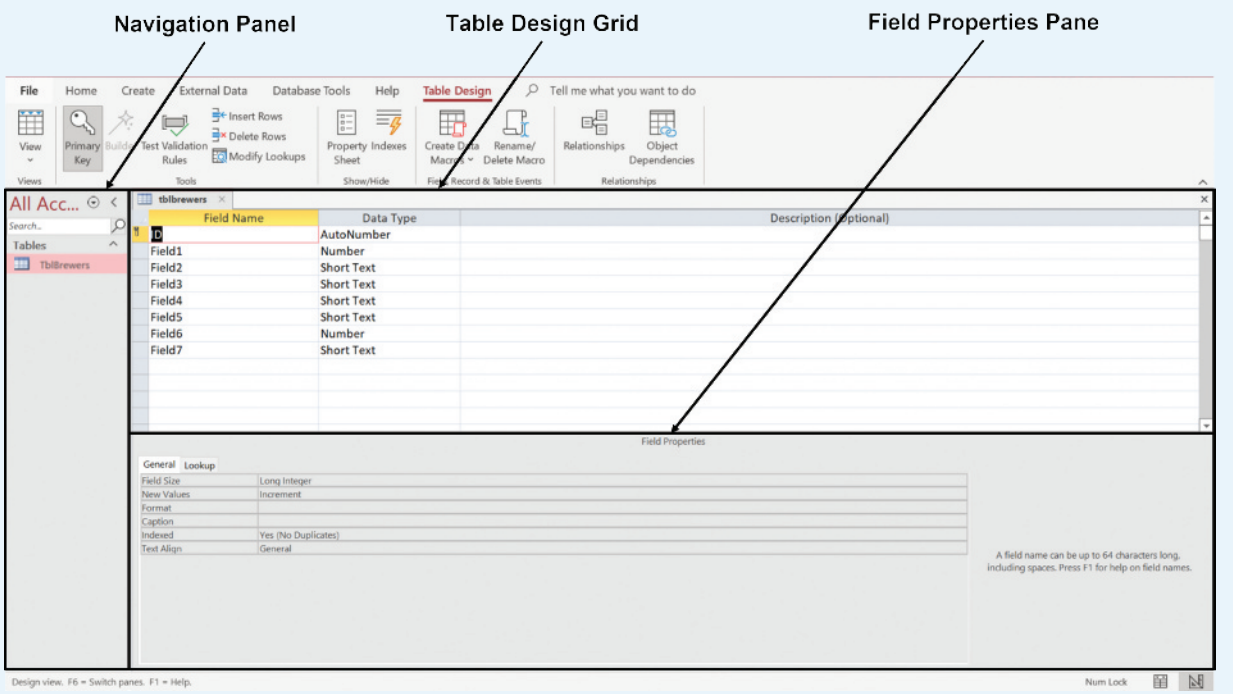
Step 2. Click the drop-down arrow in the upper right-hand corner of the selected cell

Step 3. Define the data type for the field using the drop-down menu (Figure B.8)

Notice that when you use this menu to define the data type for a field, the Field Properties Pane changes to display the properties and restrictions of the selected data type. For example, the field Brewery Name is defined as Short Text; when any row of the Table Design Grid associated with this field is selected, the Field Properties Pane shows the characteristics associated with a field of data type Short Text, including a limit of 255 characters. If we thought we might eventually do business with a brewery that has a business name that exceeds 255 characters, we may decide to select the Long Text data type for this field (Figure B.8). However, selecting a data type that allows for greater capacity will increase the size of the database and should not be used unless necessary.

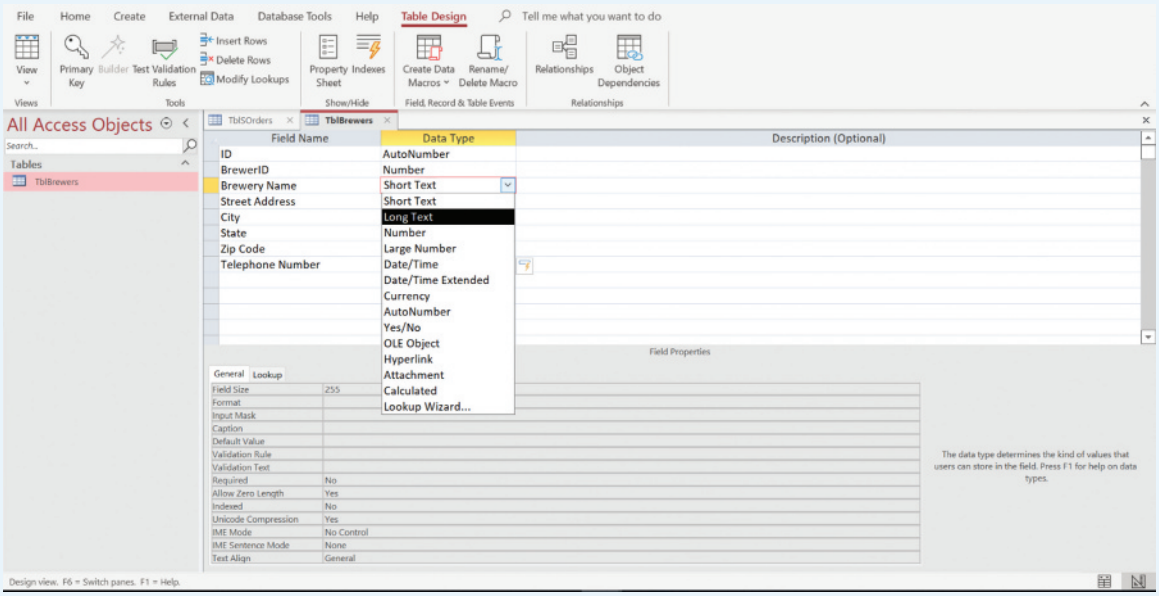
A field such as State is a good candidate for reducing the field size. If we use the official U.S. Postal Service abbreviations for the states (i.e., AL for Alabama, AK for Alaska,

Figure B.7 Design View for the Table TblBrewers



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.8 Changing the Data Type for the Brewery Name Field in the Table TblBrewers



and so on), this field would always use two characters. Note that if we violate the restriction we place on a field, Access will respond with an error statement. The restriction on length can be very helpful in this instance. Because we know that a state abbreviation is always exactly two characters, an error statement regarding the length of the State field indicates that we made an incorrect entry for this field.

Although fields such as BrewerID, Zip Code, and Phone Number are made up of numbers, we would not consider doing arithmetic operations on these cells, so we define these fields as Short Text.

After defining the data type for each of the fields to be Short Text, we can use the column labeled Description on the right side of the Table Design Grid to document the contents of each field. Here we may include the following:

- Brief descriptions of the fields (especially if our field names are not particularly meaningful or descriptive)
- Instructions for entering data into the fields (e.g., we may indicate that a telephone number is entered in the format (XXX) XXX-XXXX)
- Indications of whether a field acts as a primary or a foreign key

To change the primary key from the default field ID to BrewerID, we use the following steps:

Step 1. Click any cell in the BrewerID row
Step 2. Click the **Table Design** tab in the Ribbon
Step 3. Click the **Primary Key** icon in the **Tools** group

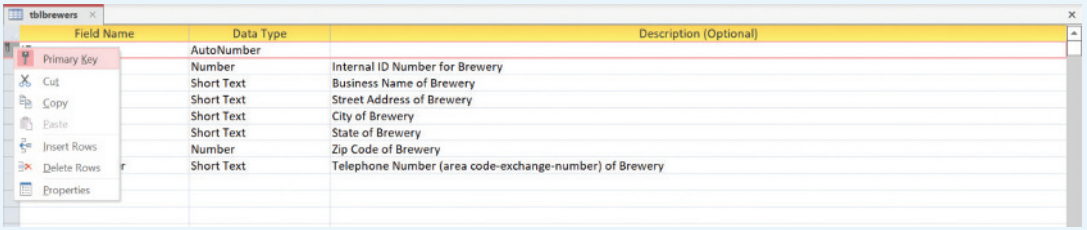
This changes the primary key from the ID field to the BrewerID field. We can now delete the ID field because it is no longer needed.

Step 4. Right-click any cell in the ID row and click **Delete Rows** (Figure B.9)
Click **Yes** when the dialog box appears to confirm that you want to delete this row

We have now created the table TblBrewers by entering the data in Datasheet view (Figure B.10) and (1) changed the name of each field, (2) identified the correct data type for each field, (3) revised properties for some fields, (4) added a description for each field, and (5) changed the primary key field to BrewerID in Design view. Alternatively, we could

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

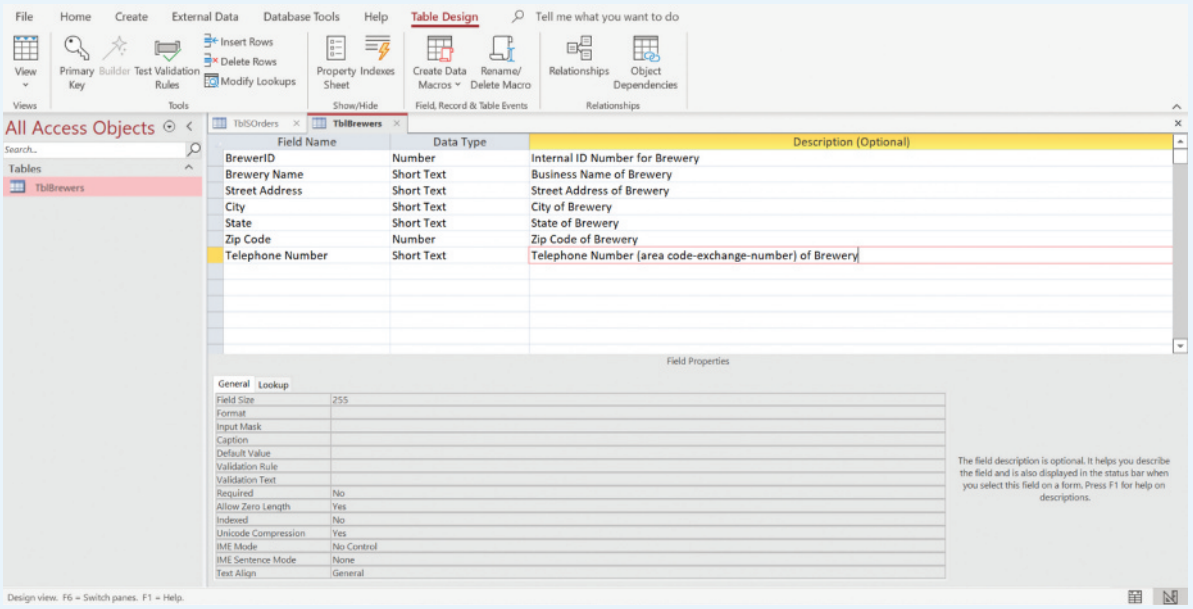
Figure B.9 Drop-Down Menu for Deleting Fields in the Design View



create a table in Design view. We first enter the field names, data types, and descriptions in the Table Design Grid. After saving this table as TblSOrders, we then move to the Database Window, which now has defined fields, and enter the information in the appropriate cells. Suppose we take this approach to create the table TblSOrders, which contains information on orders Stinson's places with the microbreweries. We have the following data for orders from the past week (Table B.2) that we will use to initially populate this table (new orders will be added to the table as they are placed).

The fields represent Stinson's internal number assigned to each order placed with a brewery (SOrderNumber), Stinson's internal identification number assigned to the microbrewery with which the order has been placed (BrewerID), the date on which Stinson's placed the order (Date of SOrder), the identification number of the Stinson's employee who placed the order (EmployeeID), an indication of whether the order was for kegs or cases of beer (Keg or Case?), and the quantity (in units) ordered (SQuantity Ordered). As before, we enter the information into the Field Name, Data Type, and Description columns of the Table Design Grid, remove the ID field, change the primary key field (this time to the field SOrderNumber), and revise the properties of the fields as necessary in the Field Properties area as shown in Figure B.11.

Figure B.10 Design View of Table Design for TblBrewers



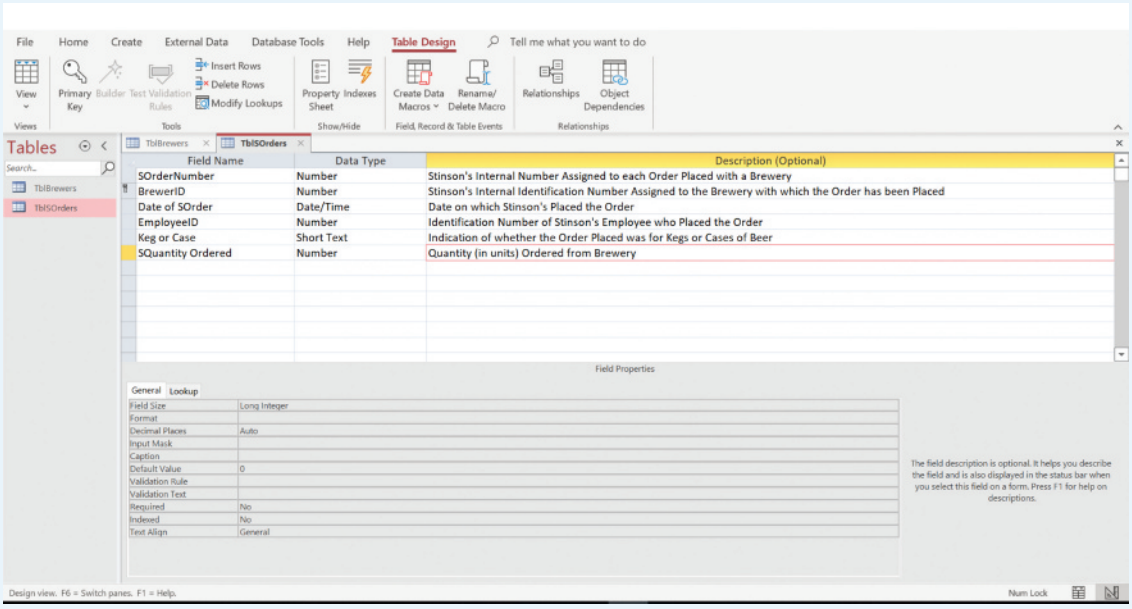
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Table B.2 Raw Data for Table TblSOrders

SOrderNumber	BrewerID	Date of SOrder	EmployeeID	Keg or Case?	SQuantity Ordered
17351	3	11/5/2023	135	Keg	3
17352	9	11/5/2023	94	Case	6
17353	7	11/5/2023	94	Keg	2
17354	3	11/6/2023	94	Keg	3
17355	2	11/6/2023	135	Keg	2
17356	6	11/6/2023	135	Case	5
17358	9	11/7/2023	94	Keg	3
17359	4	11/7/2023	135	Keg	2
17360	3	11/8/2023	94	Case	8
17361	2	11/8/2023	94	Keg	1
17362	7	11/8/2023	94	Keg	2
17363	9	11/8/2023	135	Keg	4
17364	6	11/8/2023	94	Keg	2
17365	2	11/9/2023	135	Case	5
17366	3	11/9/2023	135	Keg	4
17367	7	11/9/2023	94	Case	4
17368	9	11/9/2023	135	Keg	4
17369	4	11/9/2023	94	Keg	3

Now we return to the Database Window and manually input the data from Table B.2 into the table TblSOrders as shown in Figure B.12. Note that in both Datasheet view and Design view, we now have separate tabs with the table names TblBrewers and TblSOrders and that these two tables are listed in the Navigation Panel. We can use either Datasheet view or Design view to move between our tables.

Figure B.11 Design View of Table Design for TblSOrders



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

We can also create a table by reading information from an external file. Access is capable of reading information from several types of external files. Here we demonstrate by reading data from an Excel file into a new table TblSDeliveries. The Excel file *sdeliveries.xlsx* contains the information on deliveries received by Stinson's from various microbreweries during a recent week. The fields of this table, as shown in Figure B.12, will correspond to the column headings in the Excel worksheet displayed in Figure B.13.

The columns in Figure B.13 represent the following: Stinson's internal number assigned to each order placed with a microbrewery (SOrderNumber), Stinson's internal identification number assigned to the microbrewery with which the order has been placed (BrewerID), the identification number of the Stinson's employee who received the delivery (EmployeeID), the date on which Stinson's received the delivery (Date of SDelivery), and the quantity (in units) received in the delivery (SQuantity Delivered). To import these data directly into the table TblSDeliveries, we follow these steps:

If the Excel worksheet from which we are importing the data does not contain column headings, Access will assign dummy names to the fields that can later be changed in the Table Design Grid of Design view.

Step 1. Click the **External Data** tab in the Ribbon

Step 2. Click the **New Data Source** button, select **From File**, then click the **Excel** icon (Figure B.14)

Step 3. When the **Get External Data—Excel Spreadsheet** dialog box appears (Figure B.15), click the **Browse…** button

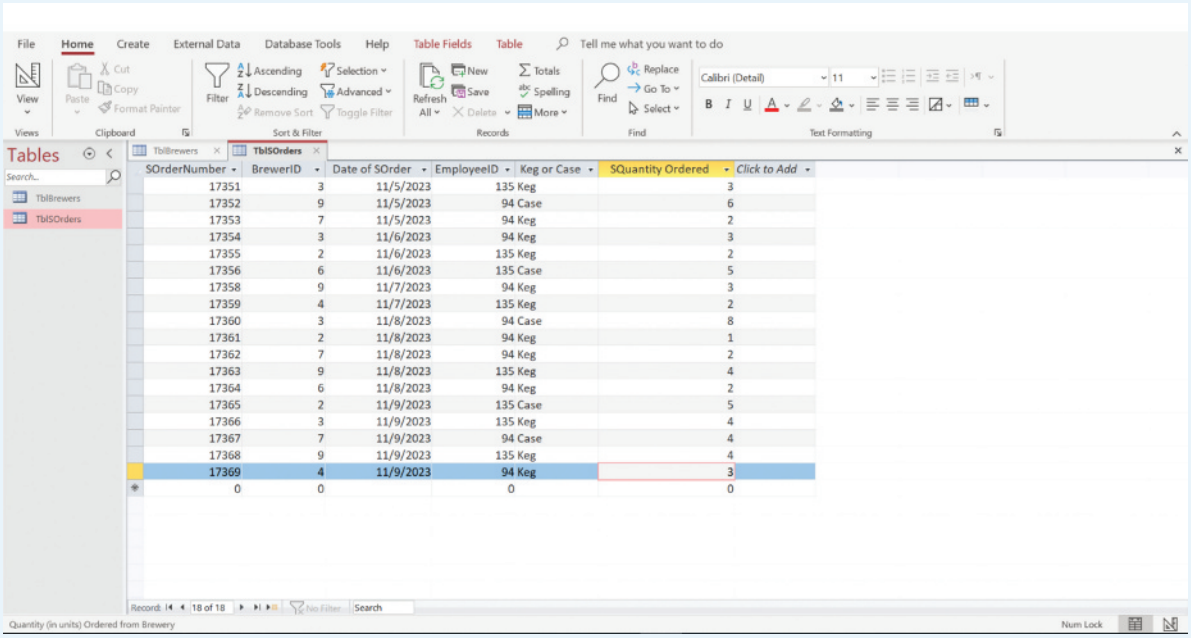
Navigate to the location of the Excel file to be imported into Access (in this case, *sdeliveries.xlsx*), and indicate the manner in which we want to import the information in this Excel file by selecting the appropriate radio button (we are importing these data to a new table, TblSDeliveries, in the current database)

Step 4. Click **OK**

Step 5. When the **Import Spreadsheet Wizard** dialog box opens (Figure B.16), arrange the information as shown in Figure B.16

Verify that the check box for **First Row Contains Column Headings** is selected because the worksheet from which we are importing the data contains column headings

Figure B.12 Datasheet View for TblSOrders



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.13 Excel Spreadsheet *sdeliveries.xlsx*

	A	B	C	D	E
1	SOrderNumber	BrewerID	EmployeeID	Date of SDelivery	SQuantity Delivered
2	17351	3	94	11/5/2023	3
3	17352	9	94	11/5/2023	6
4	17353	7	135	11/6/2023	2
5	17354	3	94	11/6/2023	3
6	17355	2	135	11/6/2023	2
7	17356	6	135	11/6/2023	5
8	17358	9	135	11/7/2023	3
9	17359	4	135	11/7/2023	2
10	17360	3	94	11/8/2023	8
11	17361	2	135	11/8/2023	1
12	17362	7	94	11/8/2023	2
13	17363	9	94	11/9/2023	4
14	17364	6	135	11/8/2023	2
15	17365	2	94	11/9/2023	5
16	17366	3	94	11/9/2023	4
17	17366	7	135	11/10/2023	4
18	17366	9	94	11/9/2023	4
19	17367	4	94	11/9/2023	3

DATA*file*
sdeliveries

Click **Next >** to open the second screen of the **Import Spreadsheet Wizard** dialog box (Figure B.17)

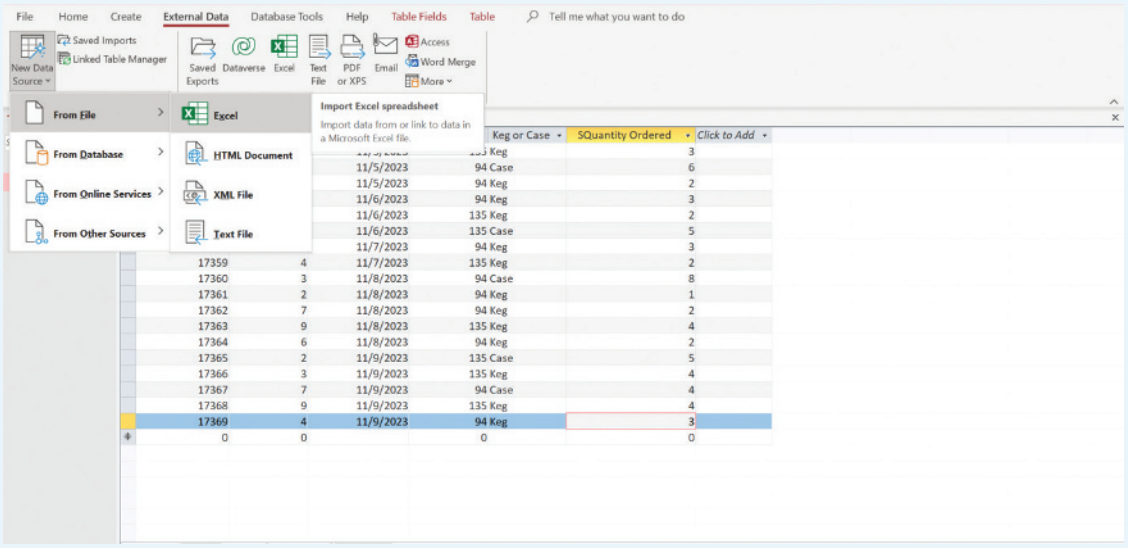
Step 6. Indicate the format for the first field (in this case, SOrderNumber) and whether this field is the primary key field for the new table (it is in this case)
Click **Next >**

If your Excel file contains multiple worksheets, you will be prompted by the Import Spreadsheet Wizard to select the worksheet from which you want to import data. After you have selected a worksheet and clicked on Next, you will automatically proceed to the screen in Figure B.16.

We continue to work through the ensuing screens of the **Import Spreadsheet Wizard** dialog box, indicating the format for each field and identifying the primary key field (SOrderNumber) for the new table. When we have completed the final screen, we click **Finish** and add the table TblSDeliveries to our database. Note that in both Datasheet view (Figure B.18) and Design view, we now have separate tabs with the table names TblBrewers, TblSOrders, and TblSDeliveries, and that these three tables are listed in the Navigation Panel.

We have now created the table TblSDeliveries by reading the information from the Excel file *sdeliveries.xlsx*, and we have entered information in the fields and identified the

Figure B.14 External Data Tab on the Access Ribbon



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.15 Get External Data—Excel Spreadsheet Dialog Box

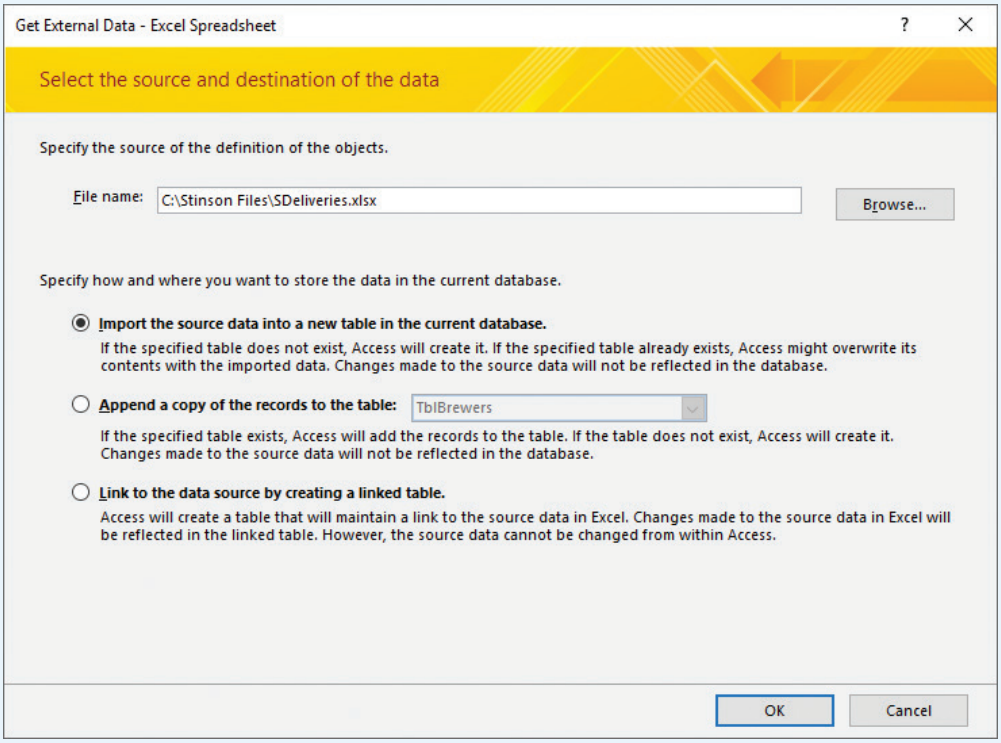
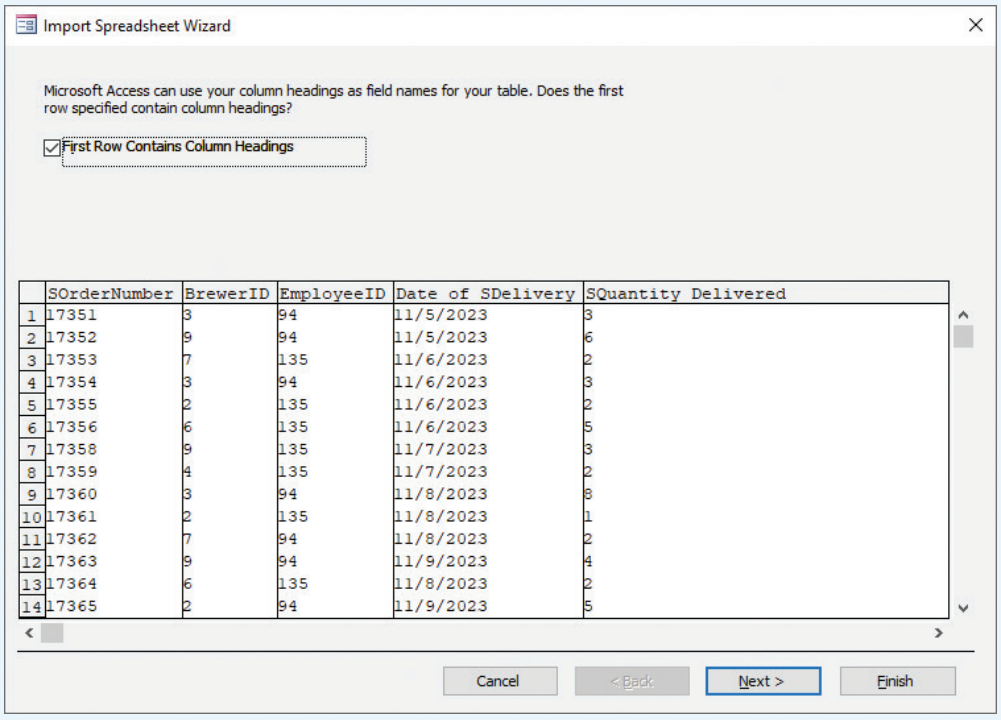


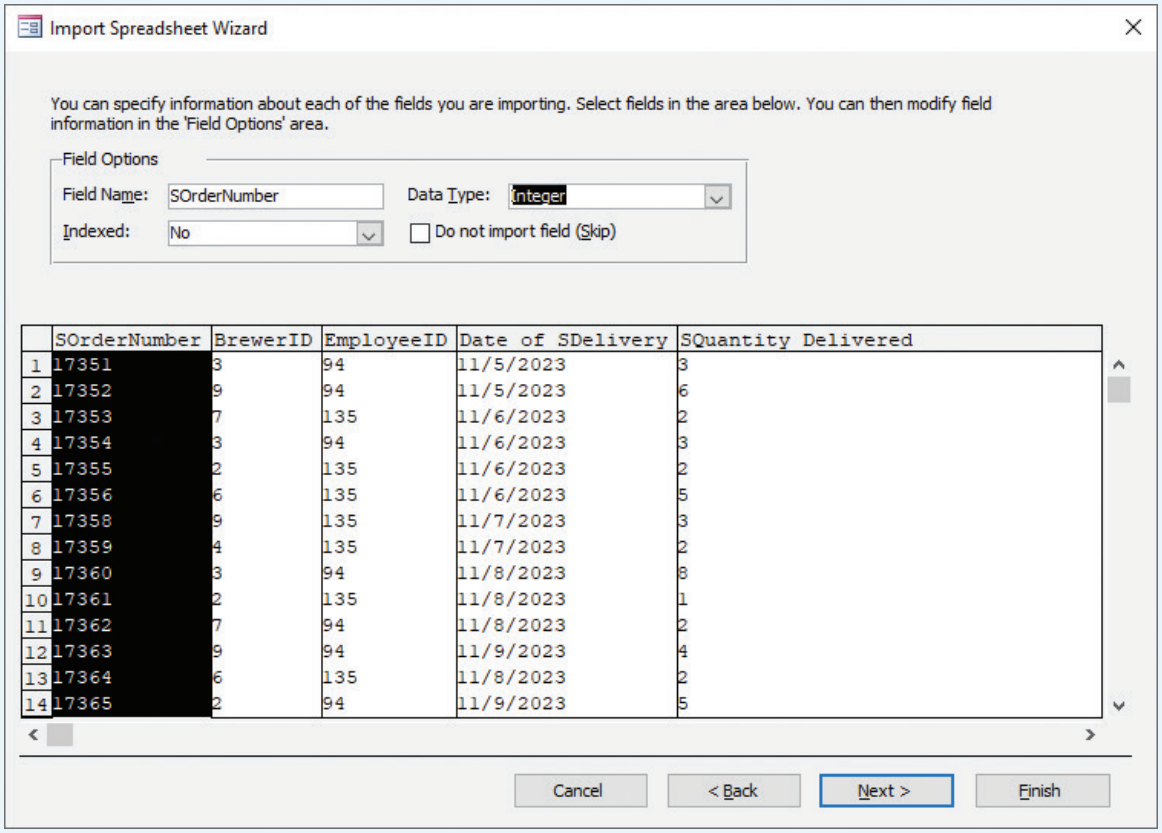
Figure B.16 First Screen of Import Spreadsheet Wizard Dialog Box



	SOrderNumber	BrewerID	EmployeeID	Date of SDelivery	SQuantity Delivered
1	17351	3	94	11/5/2023	3
2	17352	9	94	11/5/2023	6
3	17353	7	135	11/6/2023	2
4	17354	3	94	11/6/2023	3
5	17355	2	135	11/6/2023	2
6	17356	6	135	11/6/2023	5
7	17358	9	135	11/7/2023	3
8	17359	4	135	11/7/2023	2
9	17360	3	94	11/8/2023	8
10	17361	2	135	11/8/2023	1
11	17362	7	94	11/8/2023	2
12	17363	9	94	11/9/2023	4
13	17364	6	135	11/8/2023	2
14	17365	2	94	11/9/2023	5

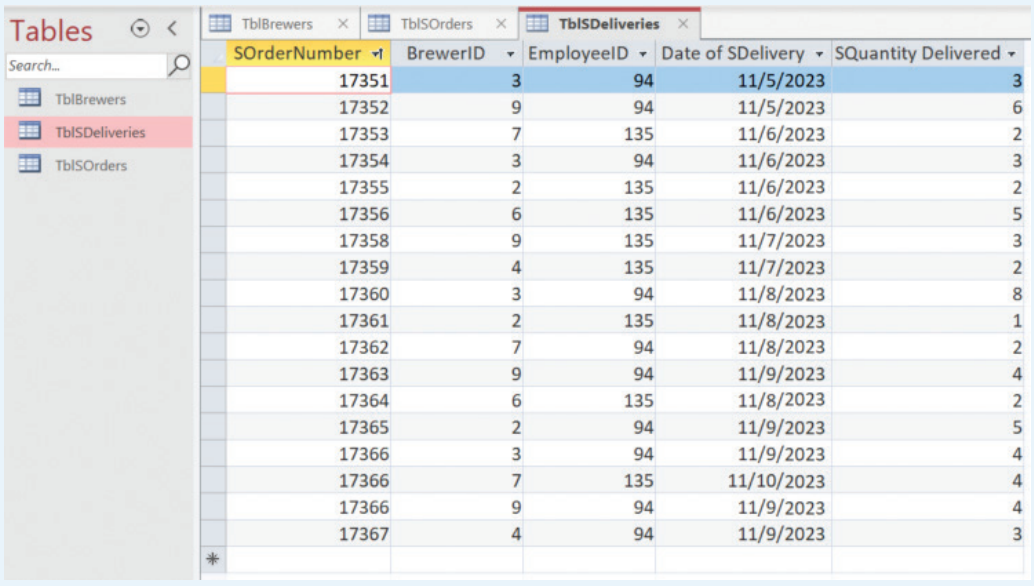
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.17 Second Screen of Import Spreadsheet Wizard Dialog Box



	SOrderNumber	BrewerID	EmployeeID	Date of SDelivery	SQuantity Delivered
1	17351	3	94	11/5/2023	3
2	17352	9	94	11/5/2023	6
3	17353	7	135	11/6/2023	2
4	17354	3	94	11/6/2023	3
5	17355	2	135	11/6/2023	2
6	17356	6	135	11/6/2023	5
7	17358	9	135	11/7/2023	3
8	17359	4	135	11/7/2023	2
9	17360	3	94	11/8/2023	8
10	17361	2	135	11/8/2023	1
11	17362	7	94	11/8/2023	2
12	17363	9	94	11/9/2023	4
13	17364	6	135	11/8/2023	2
14	17365	2	94	11/9/2023	5

Figure B.18 Datasheet View for TblSDeliveries



SOrderNumber	BrewerID	EmployeeID	Date of SDelivery	SQuantity Delivered
17351	3	94	11/5/2023	3
17352	9	94	11/5/2023	6
17353	7	135	11/6/2023	2
17354	3	94	11/6/2023	3
17355	2	135	11/6/2023	2
17356	6	135	11/6/2023	5
17358	9	135	11/7/2023	3
17359	4	135	11/7/2023	2
17360	3	94	11/8/2023	8
17361	2	135	11/8/2023	1
17362	7	94	11/8/2023	2
17363	9	94	11/9/2023	4
17364	6	135	11/8/2023	2
17365	2	94	11/9/2023	5
17366	3	94	11/9/2023	4
17366	7	135	11/10/2023	4
17366	9	94	11/9/2023	4
17367	4	94	11/9/2023	3

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

primary key field in this process. This procedure for creating a table is more convenient (and more accurate) than manually inputting the information in Datasheet view, but it requires that the data be in a file that can be imported into Access.

B.2 Creating Relationships Between Tables in Microsoft Access

One of the advantages of a database over a spreadsheet is the economy of data storage and maintenance. Information that is associated with several records can be placed in a separate table. As an example, consider that the microbreweries that supply beer to Stinson's are each associated with multiple orders for beer that have been placed by Stinson's. In this case, the names and addresses of the microbreweries do not have to be included in records of Stinson's orders, saving a great deal of time and effort in data entry and maintenance. However, the two tables, in this case the table with the information on Stinson's orders for beer and the table with information on the microbreweries that supply Stinson's with beer, must be joined (i.e., have a defined relationship) by a common field. To use the data from these two tables for a common purpose, a relationship between the two tables must be created to allow one table to share information with the other.

The first step in deciding how to join related tables is to decide what type of relationship you need to create between tables. Next we briefly summarize the three types of relationships that can exist between two tables.

One-to-Many This relationship occurs between two tables, which we will label as Table A and Table B, when the value in the common field for a record in Table A can match the value in the common field for multiple records in Table B, but a value in the common field for a record in Table B can match the value in the common field for at most a single record in Table A. Consider TblBrewers and TblSOrders with the common field BrewerID. In TblBrewers, each unique value of BrewerID is associated with a single record that contains contact information for a single brewer, while in TblSOrders each unique value of BrewerID may be associated with several records that contain information on various orders placed by Stinson's with a specific brewer. When these tables are linked through the common field BrewerID, each record in TblBrewers can potentially be matched with multiple records of orders in TblSOrders, but each record in TblSOrders can be matched with only one record in TblBrewers. This makes sense, as a single brewer can be matched to several orders, but each order can be matched to only a single brewer.

One-to-many relationships are the most common type of relationship between two tables in a relational database; these relationships are sometimes abbreviated as 1:∞.

One-to-One This relationship occurs when the value in the common field for a record in Table A can match the value in the common field for at most one record in Table B, and a value in the common field for a record in Table B can match the value in the common field for at most a single record in Table A. Here we consider TblBrewers and TblPurchasePrices, which also share the common field BrewerID. In TblBrewers, each unique value of BrewerID is associated with a single record that contains contact information for a single brewer, while in TblPurchasePrices each unique value of BrewerID is associated with a single record that contains information on prices charged to Stinson's by a specific brewer for kegs and cases of beer. When these tables are linked through the common field BrewerID, each record in TblBrewers can be matched to at most a single record of prices in TblPurchasePrices, and each record in TblPurchasePrices can be matched with no more than one record in TblBrewers. This makes sense, as a single brewer can be matched only to the prices it charges, and a specific set of prices can be matched only to a single brewer.

One-to-one relationships are the least common form of relationship between two tables in a relational database because it is often possible to include these data in a single table; these relationships are sometimes abbreviated as 1:1.

Many-to-Many This occurs when a value in the common field for a record in Table A can match the value in the common field for multiple records in Table B, and a value in the common field for a record in Table B can match the value in the common field for several records in Table A. Many-to-many relationships are not directly supported by Access but can be facilitated by creating a third table, called an *associate table*, that contains a primary key and a foreign key to each of the original tables. This ultimately results in one-to-many

Many-to-many relationships are sometimes abbreviated as ∞:∞.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

relationships between the associate table and the two original tables. Our design for Stinson's database does not include any many-to-many relationships.

To create any of these three types of relationships between two tables, we must satisfy the rules of integrity. Recall that the primary key field for a table is a field that has (and will have throughout the life of the database) a unique value for each record. Defining a primary key field for a table ensures that the table will have **entity integrity**, which means that the table will have no duplicate records.

Note that when the primary key field for one table is a foreign key field in another table, it is possible for a value of this field to occur several times in the table for which it is a foreign key field. For example, Job ID is the primary key field in the table TblJobTitle and will have a unique value for each record in this table. But Job ID is a foreign field in the table TblEmployHist, so a value of Job ID can occur several times in TblEmployHist.

Referential integrity is the rule that establishes the relationship between two tables. For referential integrity to be established, when the foreign key field in one table (say, Table B) and the primary key field in the other table (say, Table A) are matched, each value that occurs in the foreign key field in Table B must also occur in the primary key field in Table A. For instance, to preserve referential integrity for the relationship between TblEmployHist and TblJobTitle, each employee record in TblEmployHist must have a value in the Job ID field that exactly matches a value of the Job ID field in TblJobTitle. If a record in TblEmployHist has a value for the foreign key field (Job ID) that does not occur in the primary key field (Job ID) of TblJobTitle, the record is said to be **orphaned** (in this case, this would occur if we had an employee who has been assigned a job that does not exist in our database). An orphaned record would be lost in any table that results from joining TblJobTitle and TblEmployHist. Enforcing referential integrity through Access prevents records from becoming orphaned and lost when tables are joined.

Violations of referential integrity lead to inconsistent data, which results in meaningless and potentially misleading analyses. Enforcement of referential integrity is critical not only for ensuring the quality of the information in the database but also for ensuring the validity of all conclusions based on these data.

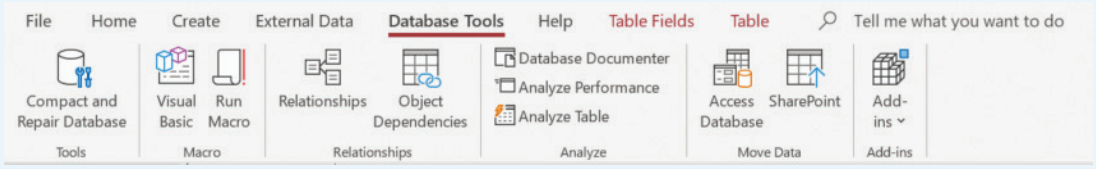
We are now ready to establish relationships between tables in our database. We will first establish a relationship between the tables TblBrewers and TblSOrders. To establish a relationship between these two tables, take the following steps:

Step 1. Click the **Database Tools** tab in the Ribbon (Figure B.19)

Step 2. Click the **Relationships** icon in the **Relationships** group
This will open the contextual tab **Relationships Design** in the Ribbon and a new display with a tab labeled **Relationships** in the workspace, as shown in Figure B.20.

Step 3 Click the **Add Tables** button in the **Relationships** group to open the **Add Tables** pane shown in Figure B.21

Figure B.19 Database Tools Tab in the Access Ribbon



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.20 Relationship Tools Contextual Tab in the Access Ribbon and Tab Labeled Relationships in the Workspace

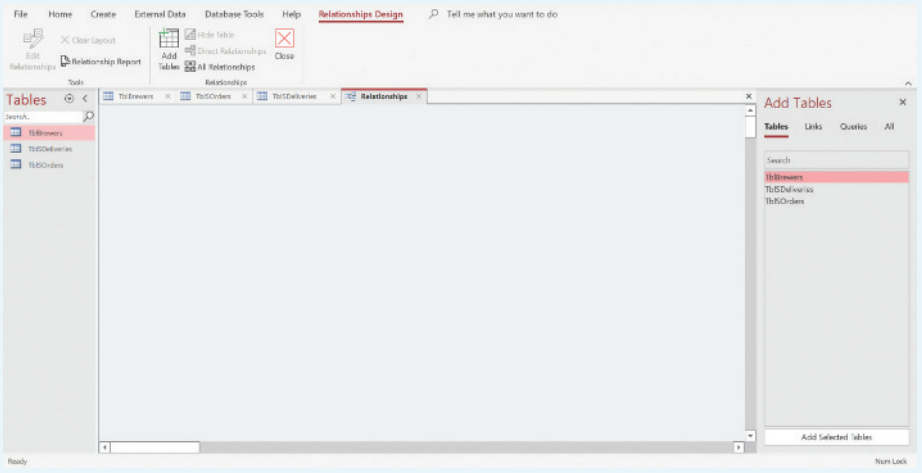
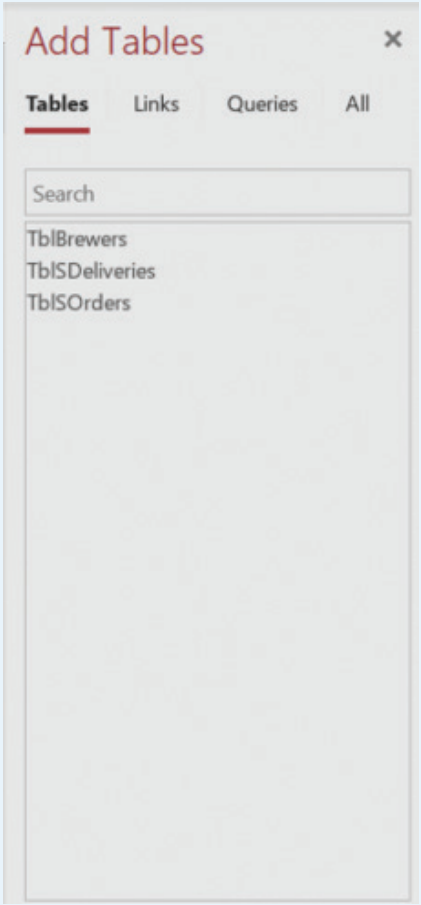
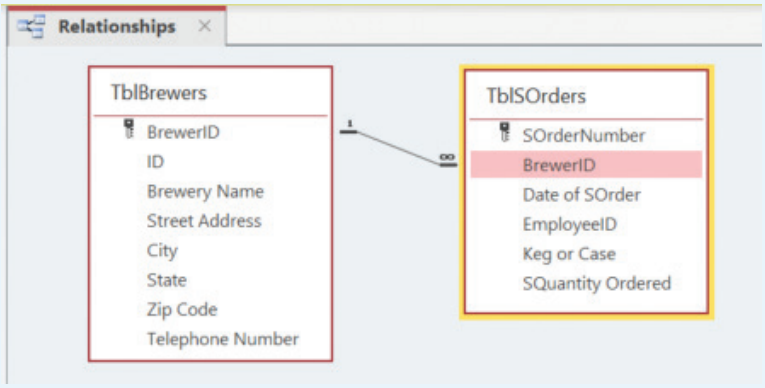


Figure B.21 Add Tables Pane



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.22 Workspace Showing the Relationship Between TblBrewers and TblSOrders



*You can select multiple tables in the **Show Table** dialog box by holding down the **Ctrl** key and selecting multiple tables.*

Step 4 Double-click **TblBrewers** in the **Add Tables** pane to open a box listing all fields for this table in the **Relationships** workspace, as shown in Figure B.22

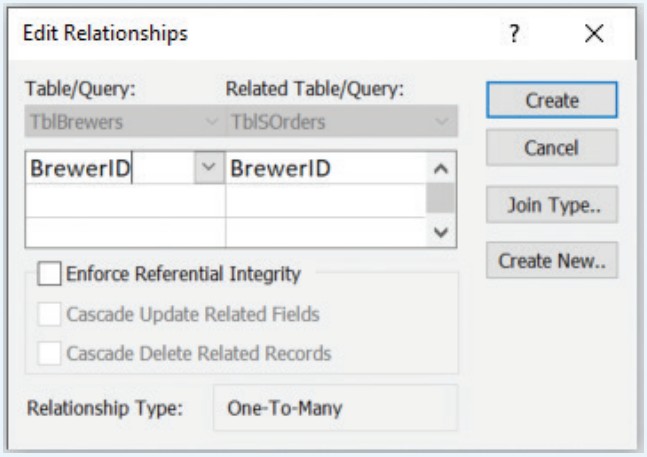
Step 5 Double-click **TblSOrders** in the **Add Tables** pane to open a box listing all fields for this table in the **Relationships** workspace, as shown in Figure B.22

Step 6 In the **Relationships** workspace, drag the **BrewerID** field in the **TblBrewers** box to the **BrewerID** field in the **TblSOrders** box to establish the relationship between these two tables based in the common field **BrewerID** (as shown in Figure B.22) and open the **Edit Relationships** dialog box (shown in Figure B.23)

Step 7 Select **Enforce Referential Integrity**

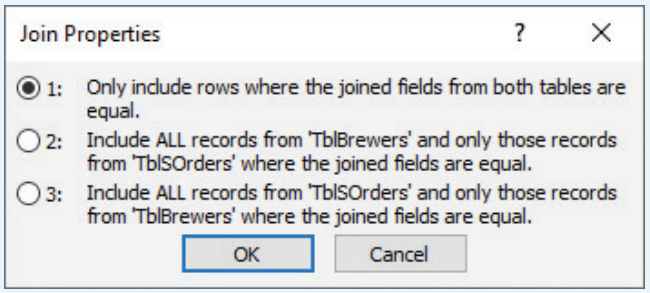
Click the **Create** button to create a one-to-many relationship between the TblBrewers table and the TblsOrders table

Figure B.23 Edit Relationships Dialog Box



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.24 Join Properties Dialog Box



By selecting the **Enforce Referential Integrity** option in the **Edit Relationships** dialog box, we can indicate that we want Access to monitor this relationship to ensure that it satisfies relational integrity. This means that every unique value in the BrewerID field in TblSOrders also appears in the BrewerID field of TblBrewers; that is, there is a one-to-many relationship between TblBrewers and TblSOrders, and Access will revise the display of the relationship, as shown in Figure B.22, to reflect that this is a one-to-many relationship.

Finally, we can click **Join Type..** in the **Edit Relationships** dialog box to open the **Join Properties** dialog box (Figure B.24). This dialog box allows us to specify which records are retained when the two tables are joined.

Once we have established a relationship between two tables, we can create new Access objects (tables, queries, reports, etc.) using information from both of the joined tables simultaneously. Suppose Stinson's will need to combine information from TblBrewers, TblSOrders, and TblSDeliveries. Using the same steps, we can also establish relationships among the three tables TblBrewers, TblSOrders, and TblSDeliveries, as shown in Figure B.25. Note that for each relationship shown in this figure, we have used the Enforce Referential Integrity option in the Edit Relationships dialog box to indicate that we want Access to monitor these relationships to ensure that they satisfy relational integrity. Thus, each relationship is identified in this case as a one-to-many relationship.

This set of relationships will also allow us to combine information from all three tables and create new Access objects (tables, queries, reports, etc.) using information from the three joined tables simultaneously.

B.3 Sorting and Filtering Records

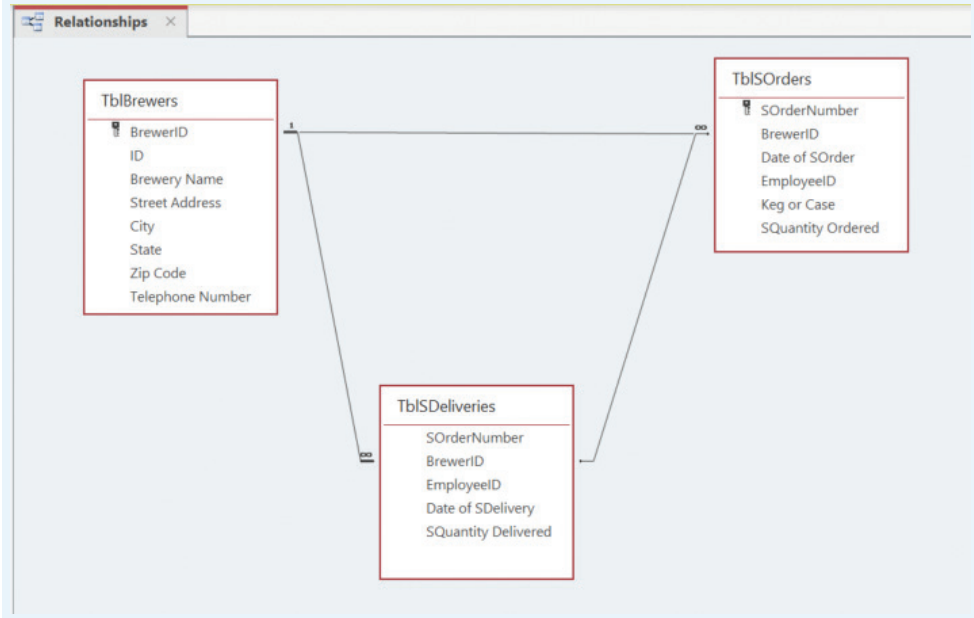
As our tables inevitably grow or are joined to form larger tables, the number of records can become overwhelming. One of the strengths of relational database software such as Access is that they provide tools, such as sorting and filtering, for dealing with large quantities of data. Access provides several tools for sorting the records in a table into a desired sequence and filtering the records in a table to generate a subset of your data that meets specific criteria. We begin by considering sorting the records in a table to improve the organization of the data and increase the value of information in the table by making it easier to find records with specific characteristics. Access allows for records to be sorted on values of one or more fields, called the *sort fields*, in either ascending or descending order. To sort on a single field, we click the Filter Arrow in the field on which we wish to sort.

Note that different data types have different sort options.

Suppose that Stinson's Manager of Receiving wants to review a list of all deliveries received by Stinson's, and she wants the list sorted by the Stinson's employee who received

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.25 Upper Portion of the Relationships Workspace Showing the Relationships Among TblBrewers, TblSOrders, and TblSDeliveries



the orders. To accomplish this, we first open the table TblSDeliveries in Datasheet view. We then click the **Filter Arrow** for the field EmployeeID (the sort field), as shown in Figure B.26; to sort the data in ascending order by values in the EmployeeID field, we click **Sort Smallest to Largest** (clicking on **Sort Largest to Smallest** will sort the data in descending order by values in the EmployeeID field). By using the Filter Arrows, we can sort the data in a table on values of any of the table's fields.

We can also use this pull-down menu to filter our data to generate a subset of data in a table that satisfies specific conditions. If we want to create a display of only deliveries that were received by employee 135, we would click the **Filter Arrow** next to EmployeeID, select only the check box for **135**, and click **OK** (Figure B.27).

Note that different data types have different filter options.

Filtering through the Filter Arrows is convenient if you want to retain records associated with several different values in a field. For example, if we want to generate a display of the records in the table TblSDeliveries associated with breweries with BrewerIDs 3, 4, and 9, we would click the **Filter Arrow** next to BrewerID, deselect the check boxes for **2, 4, 6**, and **7**, and click **OK**.

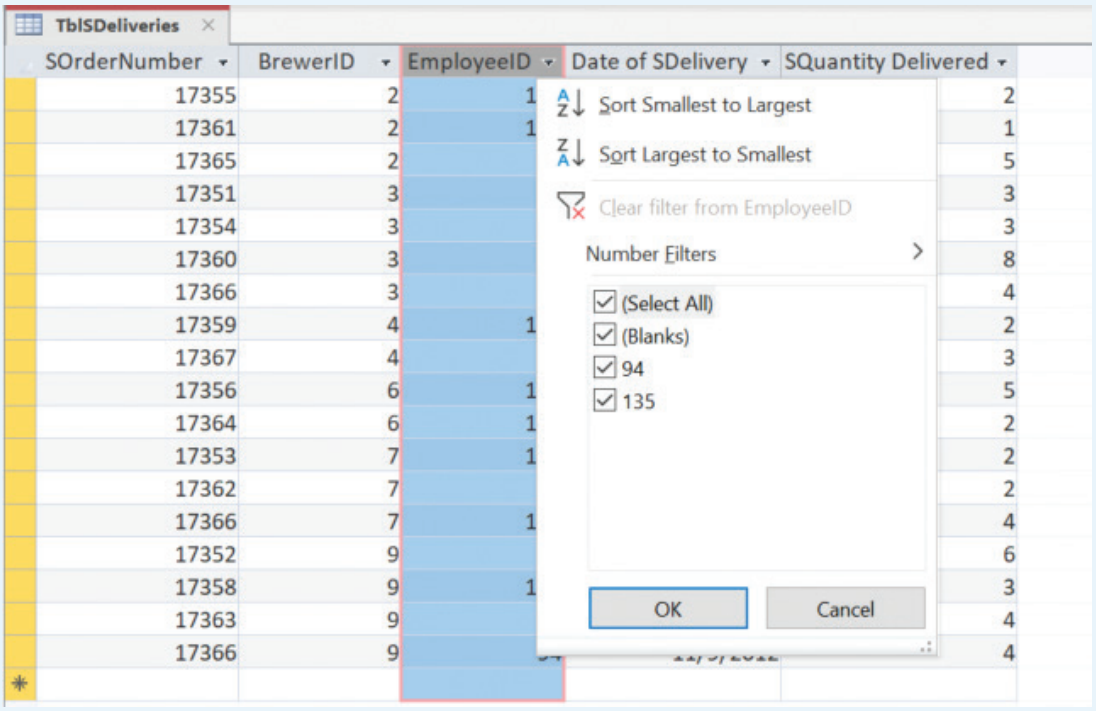
*Clicking **Selection** in the **Sort & Filter** group will also filter on values of a single field.*

The **Sort & Filter** group in the **Home** tab also provides tools for sorting and filtering records in a table. To quickly sort all records in a table on values for a field, open the table to be sorted in Datasheet view, and click any cell in the field to be sorted. Then click **Ascending** to sort records from smallest to largest values in the sort field or click **Descending** to sort records from largest to smallest in the sort field.

Access also allows for simultaneous sorting and filtering through the Advanced function in the Sort & Filter group of the Home tab; the Advanced Filter/Sort display for the table TblSDeliveries is shown in Figure B.28. Once we have opened the table to be filtered and sorted in Datasheet view, we click the **Advanced** button Advanced in the **Sort & Filter** group of the **Home** tab, as shown in Figure B.28. We then select **Advanced Filter/Sort..** From this display, we double-click the first field in the field list on which

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.26 Pull-Down Menu for Sorting and Filtering Records in a Table with the Filter Arrow



we wish to filter. The field we have selected will appear in the heading of the first column in the tabular display at the bottom of the screen. We can then indicate in the appropriate portion of this display the sorting and filtering to be done on this field. We continue this process for every field for which we want to apply a filter and/or sort, remembering that the sorting will be nested (the table will be sorted on the first sort field, and then the sort for the second sort field will be executed within each unique value of the first sort field, and so on).

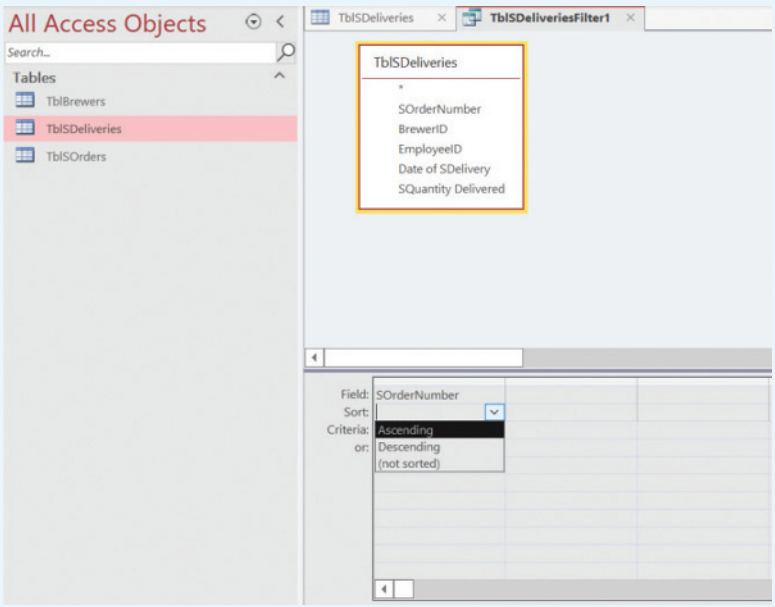
Figure B.27 Top Rows of the Tabular Display of Results of Filtering

TblSDeliveries

SOrderNumber	BrewerID	EmployeeID	Date of SDelivery	SQuantity Delivered
17353	7	135	11/6/2023	2
17355	2	135	11/6/2023	2
17356	6	135	11/6/2023	5
17358	9	135	11/7/2023	3
17359	4	135	11/7/2023	2
17361	2	135	11/8/2023	1
17364	6	135	11/8/2023	2
17366	7	135	11/10/2023	4

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.28 Advanced Filter/Sort Display for TblSDeliveries



*We can toggle between a display of the filtered/sorted data and a display of the original table by clicking on **Toggle Filter** in the **Sort & Filter** group of the **Home** tab.*

Suppose we wish to create a new tabular display of all records for deliveries from breweries with BrewerIDs of 4 or 7 for which fewer than 7 units were delivered, and we want the records in this display sorted in ascending order first on values of the field BrewerID and then on values of the field SQuantity Delivered. To execute these criteria, we perform the following steps:

Step 1. Click the **Home** tab in the Ribbon

Step 2. Click **Advanced** in the **Sort & Filter** group, and select **Advanced Filter/Sort..**

Step 3. In the **TblSDeliveries** box, double-click **BrewerID** to add this field to the first column in the lower pane of the screen
Select **Ascending** in the **Sort:** row of the **BrewerID** column in the lower pane
Enter *4* in the **Criteria:** row of the **BrewerID** column in the lower pane
Enter *7* in the **or:** row of the **BrewerID** column in the lower pane

Figure B.29 displays the lower pane of the Advanced Filter/Sort after Steps 1 through 4 have been completed.

Step 4. In the **TblSDeliveries** box, double-click **SQuantity Delivered** to add this to the second column in the lower pane of the screen
Select **Ascending** in the **Sort:** row of the **SQuantity Delivered** column in the lower pane
Enter *<7* in the **Criteria:** row of the **SQuantity Delivered** column in the lower pane

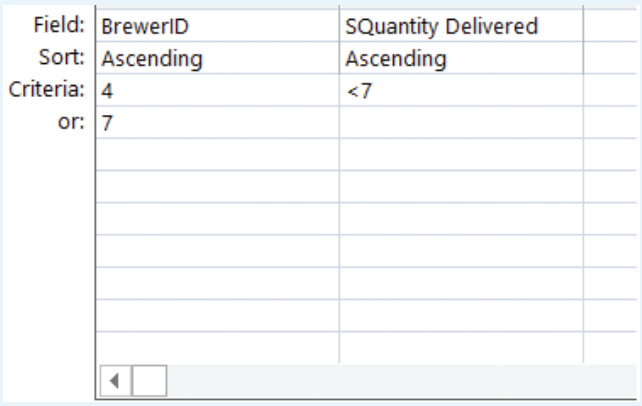
Step 5. Click **Advanced** in the **Sort & Filter** group of the **Home** tab
Click **Apply Filter/Sort**

These steps produce the tabular display shown in Figure B.30.

Note that the data, after being filtered to show only records with breweries that have values of 4 or 7 in the BrewerID field and all records with deliveries of 7 or fewer units, are sorted first in ascending order on the BrewerID field. Within each unique value in the BrewerID field, the records are sorted in ascending order on the SQuantity Delivered field.

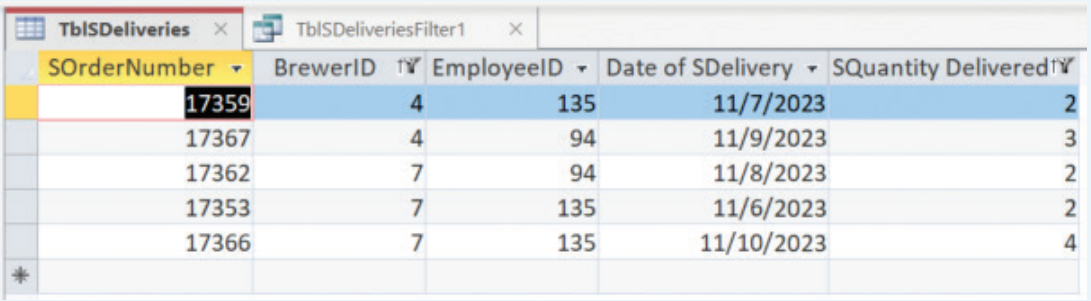
Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.29 Tabular Display of Criteria for Simultaneous Filtering and Sorting Using Advanced Filter/Sort



Field:	BrewerID	SQuantity Delivered
Sort:	Ascending	Ascending
Criteria:	4	<7
or:	7	

Figure B.30 Tabular Display of Filtered and Sorted Data Using Advanced Filter/Sort



TblSDeliveries | TblSDeliveriesFilter1

SOrderNumber	BrewerID	EmployeeID	Date of SDelivery	SQuantity Delivered
17359	4	135	11/7/2023	2
17367	4	94	11/9/2023	3
17362	7	94	11/8/2023	2
17353	7	135	11/6/2023	2
17366	7	135	11/10/2023	4

Notes + Comments

1. We can use wildcard symbols when filtering by substituting an asterisk symbol (*) for any portion of the value of a field you want to represent with a wildcard. For example, if we wanted to create a table of information on all Stinson's employees whose last names started with the letter B, we would filter the field EmpLastName in the table TblEmployees by entering *B** in the **Criteria:** row of the **Advanced Filter/Sort**. This filter will return all records that have the combination of the first letter "B" and any other following characters in the EmpLastName field.

B.4 Queries

Queries are a way of searching for and compiling data that meet specific criteria from one or more tables. They enable you to extract particular fields from a table or create a new table that combines information from several related tables.

Although there are similarities between queries and simple searches or filters, queries are far more powerful because they can be used to extract information from multiple tables.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

For example, although you could use a search in the table TblBrewers to find the name of a brewer that supplies beer to Stinson's or a filter on the table TblSOrders to view only orders placed by Stinson's for kegs of beer, neither of those approaches would let you simultaneously view both the names of brewers and the orders placed for kegs of beer. However, you could easily run a query to create a record of every order Stinson's has placed for kegs of beer that includes the name of the brewer and the corresponding order that was placed. By taking advantage of the relationships among the tables of a database, a well-designed query can yield information that would be cumbersome or difficult to discern by examining the data in individual tables.

Access allows for several types of queries. The three most commonly used are as follows:

- **Select queries**: These are the simplest and most commonly used queries; they are used to extract the subset of data from a table that satisfy one or more criteria. For example, Stinson's Manager of Receiving may want to review a list of all deliveries received by Stinson's that includes the Stinson's employee who received each order over some period of time. A select query could be applied to the table TblSDeliveries (shown in the original database design illustrated in Figure B.1) to create the subset of this table containing only the fields SOrderNumber and EmployeeID.
- **Action queries**: These queries are used to change data in existing tables. For example, the sales manager may want to increase the prices charged to retailers by Stinson's for the kegs of microbrews that Stinson's sells. The sales manager can quickly make this change through an action query applied to the table TblSalesPrices to quickly perform these calculations and modify these prices in the database. Action queries allow the user to modify many records quickly and efficiently. Access provides four types of action queries:
  - *Update* allows the values of one or more fields in the result set to be modified.
  - *Make table* creates a new table based on the results of the query.
  - *Append* is similar to a make table query, except that the results of the query are appended to an existing table.
  - *Delete* deletes all the records in the results of the query from the underlying table.
- **Crosstab queries**: These perform calculations on information in a table. Stinson's Manager of Receiving may be interested in how many kegs and cases of beer have been delivered to Stinson's and which Stinson's employee received the shipment. The manager could find this information by applying a crosstab query to the table TblSDeliveries (shown in the original database design in Figure B.1) to create a table that shows number of kegs and cases delivered by the Stinson's employee who received the shipment.

Action queries are also known as Data Manipulation Language (DML) statements.

We next review how to execute each of these types of queries in Access.

Select Queries

We start by considering the needs of Stinson's Manager of Receiving, who wants to review a list of all deliveries received by Stinson's and the Stinson's employee who received the orders during some recent week. This requires us to perform a select query on the table TblSDeliveries to create a subset of this table that includes only the fields SOrderNumber and EmployeeID for deliveries to Stinson's during the past week (the only week for which we have data in our new database) and display this subset in Datasheet view. To execute this select query, we take the following steps:

MODEL file
stinsons.accdb

Step 1. Click the **Create** tab in the Ribbon (Figure B.31)
Step 2. Click the **Query Wizard** button in the **Queries** group
Step 3. When the **New Query** dialog box appears (Figure B.32)
Select **Simple Query Wizard**
Click **OK**

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.31 Create Tab in the Access Ribbon

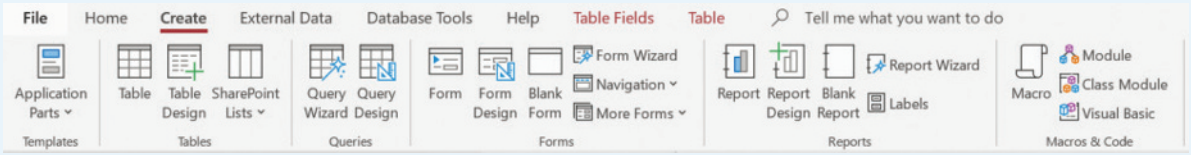
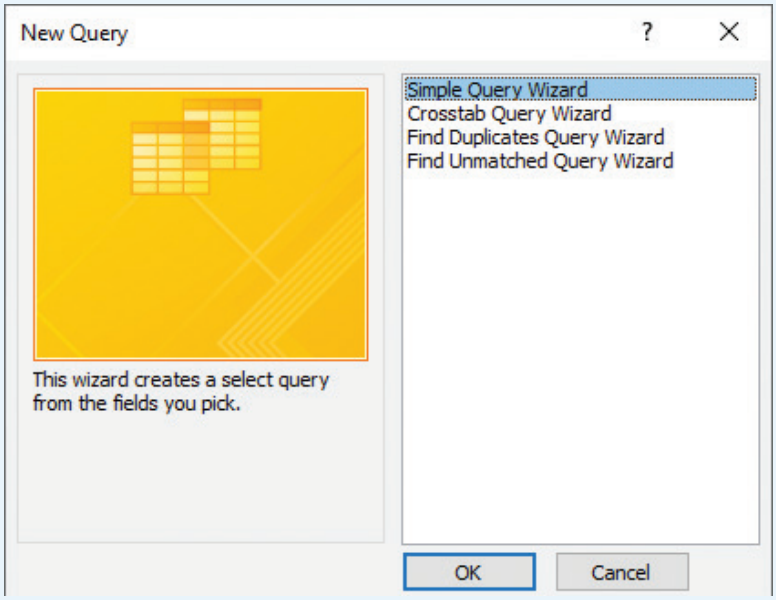


Figure B.32 New Query Dialog Box



Step 4. When the next **Simple Query Wizard** dialog box appears (see Figure B.33):
Select **Table: TblSDeliveries** in the **Tables/Queries** box
Select the fields **SOrderNumber** and **EmployeeID** from the **Available Fields:** box and move these to the **Selected Fields:** box using the > button (Figure B.33)
Click **Next >**

Step 5. When the next **Simple Query Wizard** dialog box appears (Figure B.34):
Select **Detail (shows every field of every record)**
Click **Next >**

Step 6. When the final **Simple Query Wizard** dialog box appears (Figure B.35):
Name our query by entering *TblSDeliveries Employee Query* in the **What title do you want for your query?** box
Select **Open the query to view information**
Click **Finish**

The display of the query results is provided in Figure B.36. Although Step 5 offers us the option of using the Simple Query Wizard to generate a summary display of the fields

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

A query can be saved and used repeatedly. A saved query can be modified to suit the needs of future users.

Figure B.33 First Step of the Simple Query Wizard

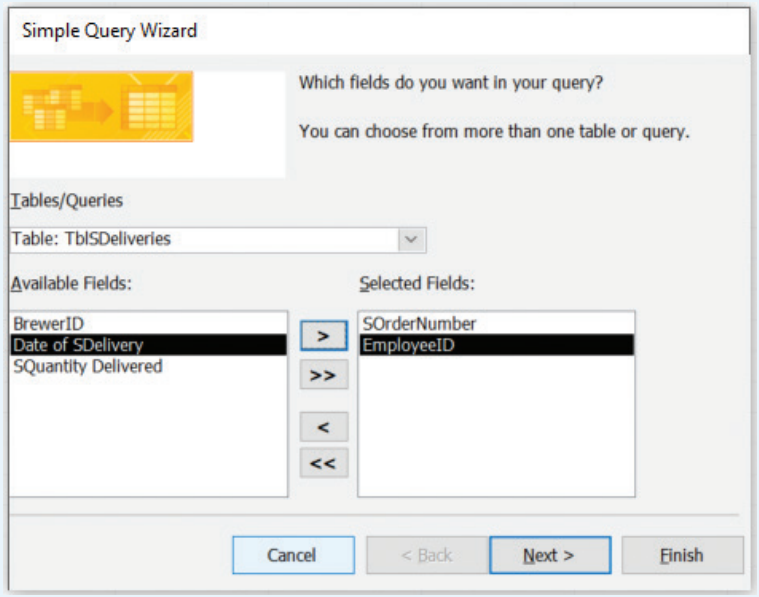
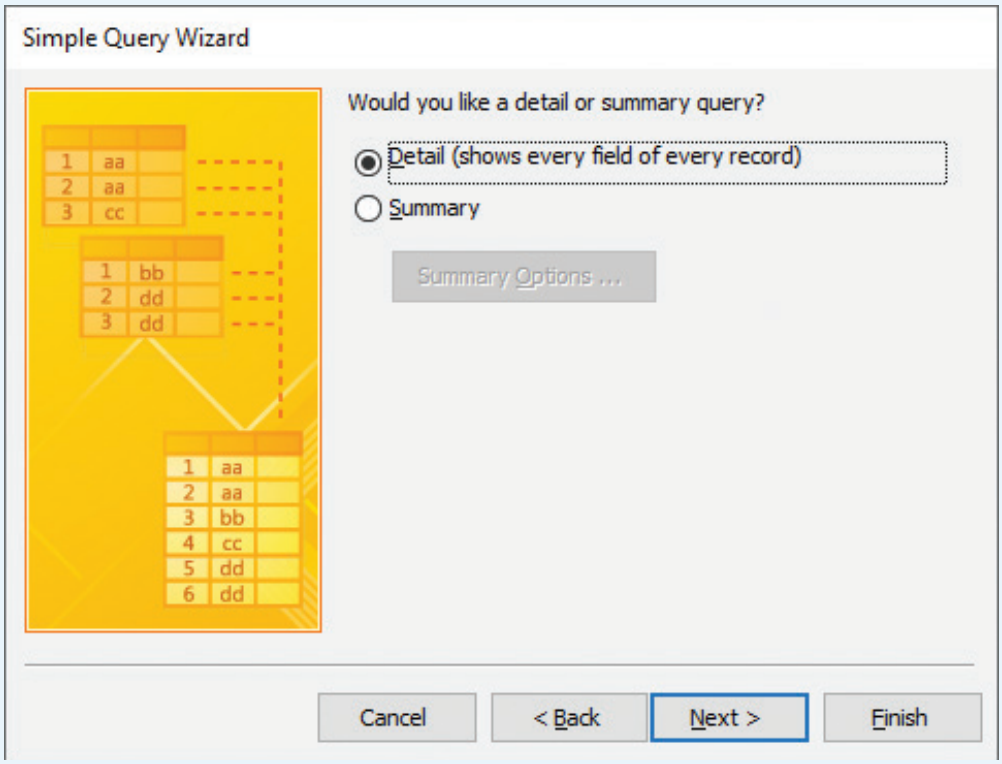


Figure B.34 Second Step of the Simple Query Wizard and the Summary Options Dialog Box



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.35 Final Step of the Simple Query Wizard

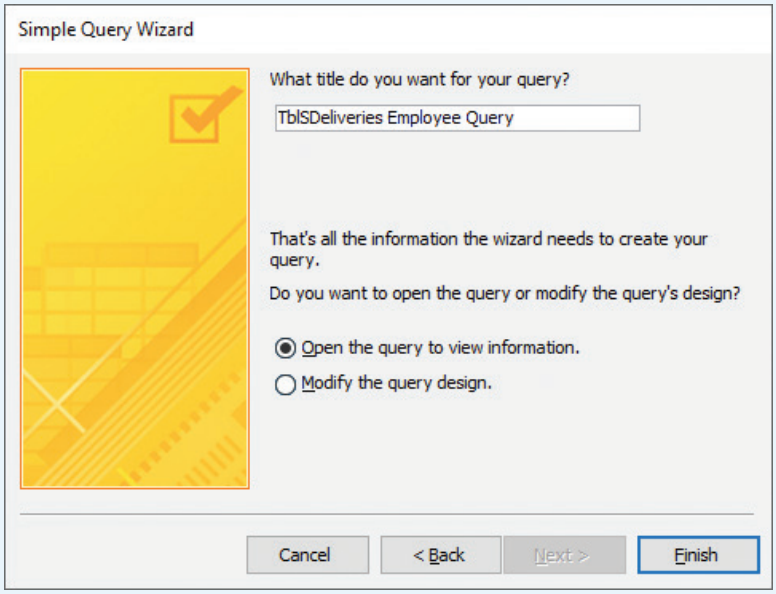
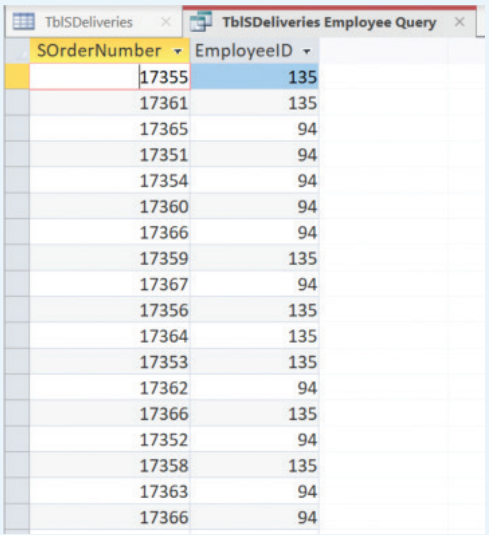


Figure B.36 Display of Results of a Simple Query



TblSDeliveries | TblSDeliveries Employee Query

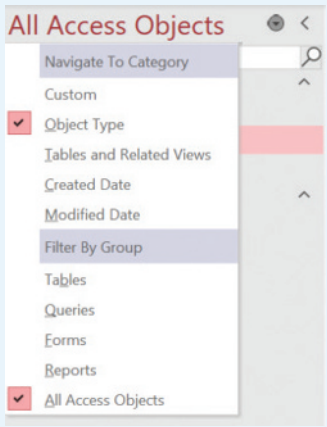
SOrderNumber	EmployeeID
17355	135
17361	135
17365	94
17351	94
17354	94
17360	94
17366	94
17359	135
17367	94
17356	135
17364	135
17353	135
17362	94
17366	135
17352	94
17358	135
17363	94
17366	94

we selected, we use the Detailed Query option here because the Manager of Receiving wants to review a list of all deliveries received by Stinson's and the Stinson's employee who received the orders during some recent week. See Figure B.34 for displays of the dialog boxes for this step of the Simple Query Wizard and Summary Options.

Note that in both Datasheet view and Design view, we now have a new tab with the table TblSDeliveries Employee Query. We can also change the Navigation Panel so that

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.37 Pull-Down Menu of Options in the Navigation Panel



it shows a list of all queries associated with this database by using the Navigation Panel's pull-down menu of options, as shown in Figure B.37.

Action Queries

Suppose that in reviewing the database system we are designing, Stinson's Sales Manager notices that we have made an error in the table TblSalesPrices. She shares with us that the price she charges for a keg of beer that has been produced by the Midwest Fiddler Crab microbrewery (value of 7 for BrewerID) should be $240, not $230 that we have entered in this table. We can use an action query applied to the table TblSalesPrices to quickly perform these changes. Because we want to modify all values of a field that meet some criteria, this is an update query. The Datasheet view of the table TblSalesPrices is provided in Figure B.38.

To make this pricing change, we take the following steps:

MODEL file
stinsons.accdb

Step 1. Click the **Create** tab in the Ribbon

Step 2. Click the **Query Design** button in the **Queries** group. This opens the **Query Design** window and the **Query Design** contextual tab (Figure B.39)

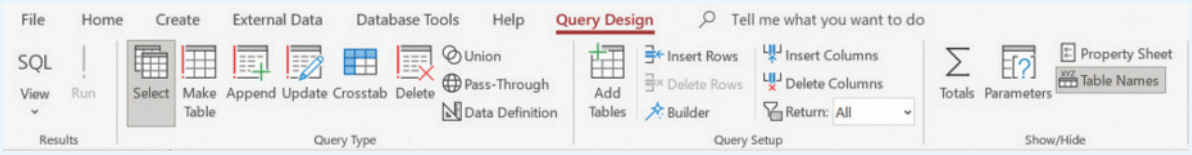
Figure B.38 Datasheet View of TblSalesPrices

TblSalesPrices

BrewerID	KegSalesPrice	CaseSalesPrice
2	$225.00	$47.00
3	$249.00	$52.00
4	$210.00	$40.00
6	$255.00	$55.00
7	$230.00	$49.00
9	$220.00	$45.00

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.39 Query Tools Contextual Tab



Step 4 produces the TblSalesPrices box that contains a list of fields in this table.

Step 3. Click **Add Tables** in the **Query Setup** group of the **Query Design** tab
Double-click **TblSalesPrices** in the **Add Tables** pane

Step 4. In the **TblSalesPrices** box, double-click **KegSalesPrice**. This opens a column labeled "KegSalesPrice" in the **Field:** row at the bottom pane of the display
Click the **Update button**, in the **Query Type** group of the **Query Design** tab
Enter *240* in the **Update To:** row of the **KegSalesPrice** column in the bottom pane of the display

Step 5. In the **TblSalesPrices** box, double-click **BrewerID** to open a second column in the bottom pane of the display labeled "BrewerID"
Enter *7* in the **Criteria:** row of the **BrewerID** column (Figure B.40)

Step 6. Click the **Run** button in the **Results** group of the **Query Design** tab
When the dialog box alerting us that we are about to update one row of the table appears, click **Yes**

Once we click **Yes** in the dialog box, the price charged to Stinson's for a keg of beer supplied by the Midwest Fiddler Crab microbrewery (BrewerID equal to 7) in the table TblSalesPrices is changed from $230.00 to $240.00.

Once saved, a query can be modified and saved again to use later.

Step 7. To save this query, click the **Save** icon in the **Quick Access** toolbar
When the **Save As** dialog box opens (Figure B.41), enter the name *Change Price per Keg Charged by a Microbrewery* for **Query Name:**
Click **OK**

Opening the table TblSalesPrices in Datasheet view (Figure B.42) shows that the price of a keg charged to Stinson's for a keg of beer supplied by the Midwest Fiddler Crab microbrewery (BrewerID equal to 7) has been revised from $230 to $240.

Crosstab Queries

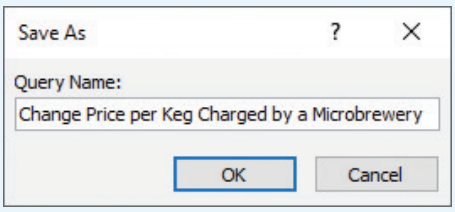
We use crosstab queries to summarize data in one field by values of one or more other fields. In our example, we will consider an issue faced by Stinson's Inventory Manager,

Figure B.40 Display of Information for the Update Query

Field:	KegSalesPrice	BrewerID
Table:	TblSalesPrices	TblSalesPrices
Update To:	240	
Criteria:		7
or:		

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.41 Save as Dialog Box



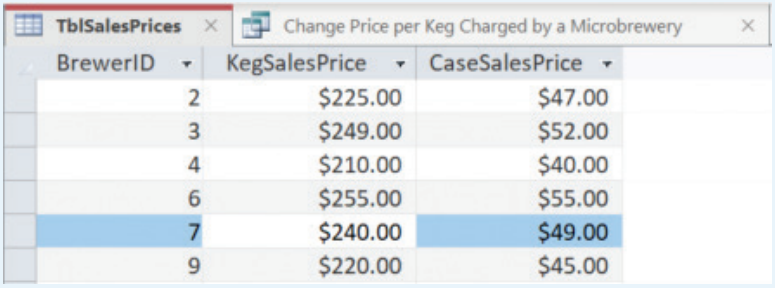
who wants to know how many kegs and cases of beer have been ordered by each Stinson's employee from each microbrewery. To provide the manager with this information, we apply a crosstab query to the table TblSOrders (shown in the original database design illustrated in Figure B.1) to create a table that shows the number of kegs and cases ordered by each Stinson's employee from each microbrewery. To create this crosstab query, we take the following steps:

MODEL file
stinsons.accdb

Step 3 produces the TblSOrders box in Access that contains a list of fields in this table.

Step 1. Click the **Create** tab in the Ribbon
Step 2. Click the **Query Design** button in the **Queries** group. This opens the **Query Design** window and the **Query Design** contextual tab
Step 3. Click the **Add Tables** button in the **Query Setup** group of the **Query Design** tab
Double-click **TblSOrders** in the **Add Tables** pane
Step 4. In the **TblSOrders** box, double-click **BrewerID**, **Keg or Case?**, and **SQuantity Ordered** to add these fields to the columns in the lower pane of the window
Step 5. In the **Query Type** group of the **Query Design** tab, click the **Crosstab button**
Step 6. In the **BrewerID** column of the window's lower pane,
Select **Row Heading** in the **Crosstab:** row
Select **Ascending** in the **Sort:** row
Step 7. In the **Keg or Case?** column of the window's lower pane,
Select **Column Heading** in the **Crosstab:** row
Select **Ascending** in the **Sort:** row

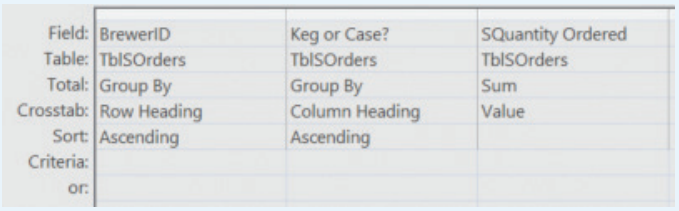
Figure B.42 TblSalesPrices in Datasheet View After Running the Update Query



BrewerID	KegSalesPrice	CaseSalesPrice
2	$225.00	$47.00
3	$249.00	$52.00
4	$210.00	$40.00
6	$255.00	$55.00
7	$240.00	$49.00
9	$220.00	$45.00

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.43 Display of Design of the Crosstab Query



Field:	BrewerID	Keg or Case?	SQuantity Ordered
Table:	TblSOrders	TblSOrders	TblSOrders
Total:	Group By	Group By	Sum
Crosstab:	Row Heading	Column Heading	Value
Sort:	Ascending	Ascending	
Criteria:			
or:			

Step 8. In the **SQuantity Ordered** column of the window's lower pane,
Select **Sum** in the **Total:** row
Select **Value** in the **Crosstab:** row

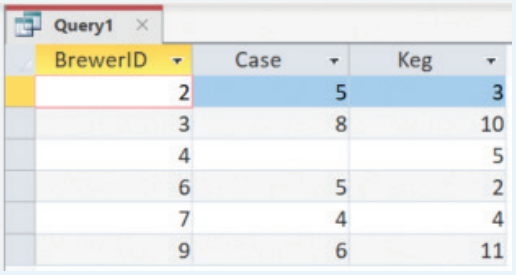
Step 9. In the **Results** group of the **Design** tab, click the **Run** button, , to execute the crosstab query

Figure B.43 displays the results of completing Steps 1 through 8 to create our crosstab query. In the first column, we have indicated that we want values of the field BrewerID to act as the row headings of our table (in ascending order), whereas in the second column we have indicated that we want values of the field Keg or Case? to act as the column headings of our table (again, in ascending order). In the third column, we have indicated that values of the field SQuantity Ordered will be summed for every combination of row (value of the field BrewerID) and column (value of the field Keg or Case?).

The results of the crosstab query appear in Figure B.44. From Figure B.44, we see that we have ordered 8 cases and 10 kegs of beer from the microbrewery with a value of 3 for the BrewerID field (the Oak Creek Brewery).

Step 10. To save the results of this query, click the **Save** icon, , in the **Quick Access** toolbar
When the **Save As** dialog box opens, enter *Brewer Orders Query* for **Query Name:**
Click **OK**

Figure B.44 Results of Crosstab Query



Query1

BrewerID	Case	Keg
2	5	3
3	8	10
4		5
6	5	2
7	4	4
9	6	11

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Notes + Comments

1. Action queries permanently change the data in a database, so we suggest that you back up the database before performing an action query. After you have reviewed the results of the action query and are satisfied that the query worked as desired, you can then save the database with the results of the action query. Some cautious users save the original database under a different name so that they can revert to the original preaction query database if they later find that the action query has had an undesirable effect on the database.
2. Crosstab queries do not permanently change the data in a database.
3. The Make Table, Append, and Delete action queries work in manners similar to Update action queries and are also useful ways to modify tables to better suit the user's needs.

B.5 Saving Data to External Files

You can open the file stinsons.accdb *and follow these steps to reproduce an external Excel file of the data in TblSOrders.*

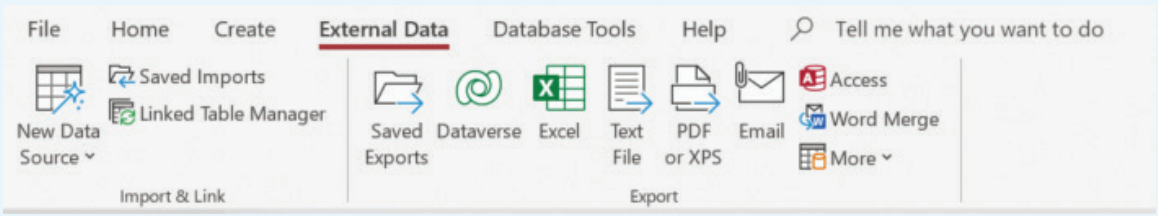
Access can export data to external files in formats that are compatible with a wide variety of software. To export the information from the table TblSOrders to an external Excel file, we take the following steps:

Step 1. Click the **External Data** tab in the Ribbon (Figure B.45)
Step 2. In the **Navigation Panel**, click **TblSOrders**
Step 3. In the **Export** group of the **External Data** tab, click the **Excel** icon,
Step 4. When the **Export—Excel Spreadsheet** dialog box opens (Figure B.46), click the **Browse…** button
Find the destination where you want to save your exported file and then click the **Save** button
Verify that the correct path and filename are listed in the **File Name**: box (*TblSOrders.xlsx* in this example)
Verify that the **File format:** is set to **Excel Workbook (*.xlsx)**
Select the check boxes for **Export data with formatting and layout**. and **Open the designation file after the export operation is complete**.
Click **OK**

After we complete Step 4, another dialog box asks us if we want to save the steps we used to export the information in this table; this can be useful if we have to export similar data again.

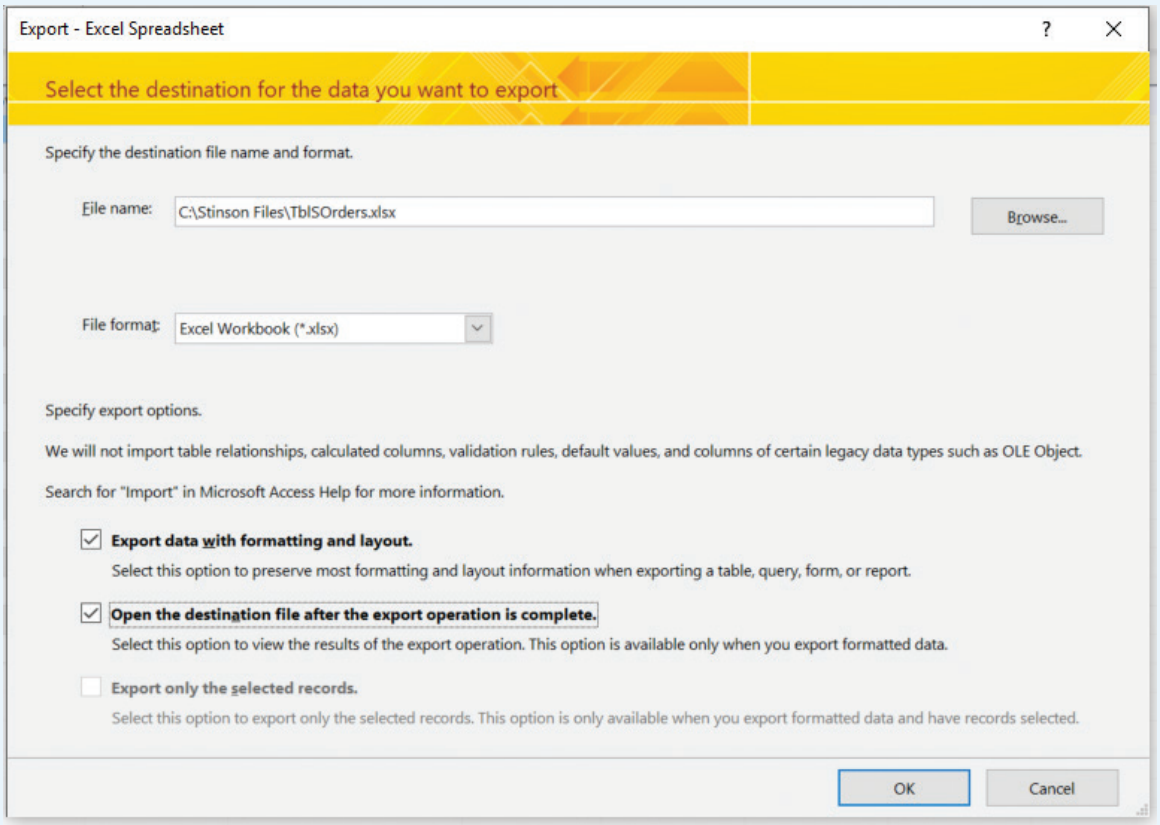
The preceding steps export the table TblSOrders from Access into an Excel file named *TblSOrders.xlsx*. Exporting information from a relational database such as Access to Excel allows one to apply the tools and techniques covered throughout this textbook to a subset of a large data set. This can be much more efficient than using Excel to clean and filter large data sets.

Figure B.45 External Data Tab in Access



Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Figure B.46 Export—Excel Spreadsheet Dialog Box



Summary

The amount of data available for analyses is increasing at a rapid rate, and this trend will not change in the foreseeable future. Furthermore, the data used by organizations to make decisions are dynamic, and they change rapidly. Thus, it is critical that a data analyst understand how data are stored, revised, updated, retrieved, and manipulated. We have reviewed tools in Microsoft Access® that can be used for these purposes.

In this appendix we have reviewed the basic concepts of database creation and management that are important to consider when using data from a database in an analysis. We have discussed several ways to create a database in Microsoft Access®, and we have demonstrated Access tools for preparing data in an existing database for analysis. These include tools for reading data from external sources into tables, creating relationships between tables, sorting and filtering records, designing and executing queries, and saving data to external files.

Glossary

Action queries Queries that are used to change data in existing tables. The four types of action queries available in Access are update, make table, append, and delete.

Compound primary key A unique identifier for each record in the table formed by combining two or more fields that is used to identify how records from several tables in a database are logically related.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Crosstab queries Queries that are used to summarize data in one field across the values of one or more other fields.
Database A collection of logically related data that can be retrieved, manipulated, and updated to meet a user's or organization's needs.
Datasheet view A view used in Access to control a database; provides access to tables, reports, queries, forms, etc. in the database that is currently open. This view can also be used to create tables for a database.
Design view A view used in Access to define or edit a database table's fields and field properties as well as to rearrange the order of the fields in the database that is currently open.
Entity integrity The rule that establishes that a table has no duplicate records. Entity integrity can be enforced by assigning a unique primary key to each record in a table.
Fields The variables or characteristics for which data have been collected from the records.
Foreign key field A field that is permitted to have multiple records with the same value.
Form An object that is created from a table to simplify the process of entering data.
Leszynski/Reddick guidelines A commonly used set of standards for naming database objects.
Many-to-many Sometimes abbreviated as ∞:∞, a relationship for which a value in the common field for a record in one table (say, Table A) can match the value in the common field for multiple records in another table (say, Table B), and a value in the common field for a record in Table B can match the value in the common field for several records in Table A.
One-to-many Sometimes abbreviated as 1:∞, a relationship between tables for which a value in the common field for a record in one table (say, Table A) can match the value in the common field for multiple records in another table (say, Table B), but a value in the common field for a record in Table B can match the value in the common field for at most a single record in Table A.
One-to-one Sometimes abbreviated as 1:1, a relationship between tables for which a value in the common field for a record in one table (say, Table A) can match the value in the common field for at most one record in another table (say, Table B), and a value in the common field for a record in Table B can match the value in the common field for at most a single record in Table A.
Orphaned A record in a table that has a value for the foreign key field of a table that does not match the value in the primary key field for any record of a related table. Enforcing referential integrity prevents the creation of orphaned records.
Primary key field A field that must have a unique value for each record in the table and is used to identify how records from several tables in a database are logically related.
Query A question posed by a user about the data in the database.
Records The individual units from which the data for a database have been collected.
Referential integrity The rule that establishes the proper relationship between two tables.
Report Output from a table or a query that has been put into a specific prespecified format.
Select queries Queries that are used to extract the subset of data that satisfy one or more criteria from a table.
Table Data arrayed in rows and columns (similar to a worksheet in an Excel spreadsheet) in which rows correspond to records and columns correspond to fields.

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Index

A

Absolute references, 653
Accessing data, 153–157
ACCO Brands Corporation case study, 472
Accounting analytics, 9–10
Accuracy, 577
Activation function, 549, 550, 602, 603
Addition law, 203–205
Additivity, linear optimization, 748
Advanced analytics, 9
African elephant populations case study, 81–82
Agencia Estatal de Seguridad Aerea (AESA), 850
All-integer linear program, 806
Alternative hypotheses, 349–352
Alternative optimal solutions
  in binary optimization, 827–829
  for linear programs, 758–759, 781–782
Alumni giving case study, linear regression, 466–467
American Statistical Association (ASA), 13–14
Analytical methods and models
  descriptive analytics, 4–5
  predictive analytics, 5
  prescriptive analytics, 5–6
Antecedent, 284
Arcs, 768
Area under the ROC curve, 582
Arithmetic mean, 46, 47
Artificial intelligence (AI), 8
Assignment problem, 776–779
Association rules, 284–287
  evaluation, 286–287
Auditing, spreadsheet models, 656–660
  Error Checking, 658, 659, 660
  Evaluate Formulas, 658, 659
  Show Formulas, 656
  Trace Dependents, 656, 657
  Trace Precedents, 656, 657
  Watch Window, 659, 660
Automated recommender systems case study, 258
Autoregressive models, 496
AVERAGE function, 218, 357, 361

B

Backpropagation, 553, 607
Backward elimination procedure, 441, 442
Bagging approach, 543–544, 545, 546, 595
Bag of words, 288
Bank location problem, 820–822
Banner ad revenue, maximizing, 772–776
Bar charts, 108–111
Base-case scenario, 673–674
Bass forecasting model, 866–869
Batch size, 553, 606
Bayes' theorem, 211–213, 910–913
Bell-shaped distribution, 57–59, 232–236
Bernoulli distribution, 737
Best-case scenario, 674
Best subsets method, 556, 609
Best subsets procedure, 441, 442
Beta distribution, random variable, 706–708, 734–735
Bias term, 549, 550, 554, 602, 603, 607, 608
Bid fraction values, 700–708
Big data
  attributes, 370
  and confidence intervals, 306–307
  defined, 6
  estimation, 369–370
  and hypothesis testing, 374–376
  implications for confidence intervals, 373–374
  overview, 6–9
  and *p* values, 374–375
  and sampling error, 371–372
  statistical inference and, 367–376
  tall data, 370
  variety, 7, 370
  velocity, 6–7, 370
  veracity, 7–9, 370
  volume, 6, 7, 370
  wide data, 370
Big Ten Conference case study, 315–316
Bimodal data, 48
Binary document-term matrix, 288
Binary integer linear program, 807
Binary variables, 269, 270
  applications, 815–825
  bank location problem, 820–822
  capital budgeting problem, 815–816
  fixed-cost problem, 816–820
  modeling flexibility provided by, 825–827
  optimization alternatives, 827–829
  product design and market share optimization, 822–825
Binding constraint, 753
BINOM.DIST function, 223
Binomial distribution, 737
Binomial probability distribution, 221–223
Bins, frequency distributions, 32
  number of, 35
  range spanned by, 36
  width of, 35–36
Boosting method, 547–548, 599
Boxplots, distribution analysis, 59–62
Branch-and-bound approach, 810
Branches, 896
Branch probabilities, computation with Bayes' theorem, 910–913
Breakpoint, nonlinear relationships, 434
Bubble charts, 113–115
Bundle pricing, 872
Business analytics
  decision making and, 894
  defined, 3–4
  methods and models, 4–6
  in practice, 9–12
  role of, 2
  spectrum of, 10
  unintended consequences of, 13

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Business analytics (*Continued*)
usage, legal and ethical issues in, 12–14
use of, 894
Business cycles, 481

C

CAP (certified analytics professional), 14
Capital budgeting problem, 815–816
Cartograms, 129–131
Categorical data, 24
frequency distributions for, 32–33
Categorical independent variables, 424–428
Categorical variables, 185–186, 496–497
Causal/exploratory forecasting methods, 473
Causal forecasting, 500–503
Causal variables, 503–504
Census, 22, 320–321
Central limit theorem, 333
Centroid linkage, 283
Chain rule, 553
Chance events, 895
Chance nodes, 896
Charts
bar, 108–111
bubble, 113–115
column, 108–111
defined, 101
Excel, 401–402
line, 92, 93, 104–108
parallel-coordinates, 126
pie, 112
PivotCharts (Excel), 117–118
scatter, 62–63, 66, 101–103
scatter-chart matrix, 115–116
stock, 124–126
table lens, 116–117
tables *vs.,* 92–93
three-dimensional, 112
waterfall, 122–124
Choropleth maps, 127–129
Cincinnati Zoo & Botanical Gardens, 84–85
Class 0 error rate, 577–582
Class 1 error rate, 577–582
Classification, 572
binary, performance measures for, 576–582
k-nearest neighbor, 587–590
with logistic regression, 582–587
neural network. *See* Neural network
probabilities, 578
Classification trees, 591–600
defined, 591
ensemble methods, 595–600
generating predictions with, 593–595
Class imbalanced data, 574–576
Cleaning, 153, 167–176
erroneous outliers/erroneous values/duplicate records, 170–176
missing data, 167–170
Cloud computing, 2, 8
Cluster analysis, 266–283
hierarchical clustering, 266, 275–283
k-means clustering, 266, 271–275
measuring distance between observations, 267–271
Clustered-column/clustered-bar chart, 110, 111
Cluster interpretability, 275
Coca-Cola, 9
Coefficient allowable increase and decrease, 764
Coefficient of determination, 405–406, 409
Coefficient of variation, 54
Cognitive load, 86
Column charts, 108–111
Complement of an event, 202–203
Complete linkage, 276
Component ordering Model, 653
CONCAT function, 39
Concave function, 856
Conditional constraint, 826–827
Conditional probability, 205–213
Bayes' theorem, 211–213, 912
independent events, 210
multiplication law, 210
Confidence, 285
Confidence coefficient, 343
Confidence interval, 448, 449, 610
big data and, 306–307
implications of big data for, 373–374
individual linear regression parameters, 419–420
statistical inference, 343–344, 363–364
Confidence level, 419–420
Confusion matrix, 577, 579
Conjoint analysis, 822–825
Consequent, 284
Conservative approach, 898
Constants, smoothing, 490
Constrained problem, nonlinear optimization models, 851–853
Constraints
binding, 753
conditional, 826–827
corequisite, 826–827
defined, 744
greater-than-or-equal-to, 763
k out of *n* alternatives, 826
less-than-or-equal-to, 778
linear optimization model, 744, 746–748
multiple-choice, 825–826
mutually exclusive, 825–826
nonnegativity, 748, 749
relaxing/loosening, 763
shadow price for, 763
tightening, 763
Consumer Research, Inc. case study, 468
Consumer surplus, 874
Continuous probability distributions, 227–239, 701, 734–736
exponential, 236–239
normal, 231–236
risk analysis, 676
triangular, 229–231
uniform, 227–229
Continuous random variables, 214–215, 227
Controllable inputs, 673
Convergence parameter, Evolutionary method, 871
Convex function, 857
Convex hull, 809
Corequisite constraint, 826–827
Corpus, 288
Correlation coefficient, 67–68
Cosine distance, 293
COUNTA function, 347, 366
COUNTBLANK function, 168
COUNT function, 347, 357, 361

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

COUNTIF function, 42–43, 347, 366
COUNTIFS function, 175
Covariance, 64–66
COVARIANCE.S function, 64, 65
Coverage error, 368
Cross-sectional data, 24–25
Crosstabulation, 94–96
Cross-validation, 443
csv file, 154
Cumulative frequency distribution, 43–45
Cumulative probabilities, 231
Current Population Survey (CPS), 22
Custom discrete distribution, 737, 738
Custom discrete probability distribution, 215–217
Cutoff value, 578, 579
Cutting plane approach, 810
Cyclical patterns, time series analysis, 481

D

Data
  accessing, 153–157
  arrangement of, 159–161
  bimodal, 48
  categorical, 24
  cross-sectional, 24–25
  defined, 23
  distributions from, 32–45
    cumulative, 43–45
    frequency distributions for categorical data, 32–33
    frequency distributions for quantitative data, 35–37
    frequency polygons, 42–43
    histograms, 37–41
    relative and percent frequency distributions, 33–35
  enriching, 182–186
  in Excel
    conditional formatting, 30–32
    sorting and filtering, 27–30
  hierarchical, 121
  missing, 167–170
  multimodal, 48
  numeric, 159
  overview of using, 23–24
  quantitative, 24
  raw, 152, 157–158
  semi-structured, 158
  sources of, 25–26
  stacked, 161
  structured, 157
  subsetting, 177–179
  supplementing, 179–182
  tall, 370
  text, 159
  time series, 24–25
  types of, 24–27
  unstacked, 161
  unstructured, 157, 287
  wide, 370
Data dashboards, 131–133
  applications of, 132–133
  defined, 4
  principles of effective, 131–132
Data-driven decision making, 2
Data formatting, 159
Data-ink ratio, 89–91
Data mining
  defined, 4
  descriptive, 257–317
    association rules, 284–287
    automated recommender systems case study, 258
    Big Ten Conference case study, 315–316
    cluster analysis. *See* Cluster analysis
    dimension reduction, 259–266
    Know Thy Customer (KTC) case study, 316–317
    text mining, 287–295
  Grey Code Corporation (GCC) case study, 630–632
  Orbitz case study, 524
  predictive, 523–632
    classification trees, 591–600
    classification with logistic regression, 582–587
    data partitioning, 524, 526–535, 572, 573–576
    data preparation, 524, 526–535, 572, 573–576
    data sampling, 524, 526–535, 572, 573–576
    defined, 524
    feature selection, 555–558, 609–611
    housing bubble case study, 568–569
    k-nearest neighbor classification, 587–590
    k-nearest neighbor regression, 535–538
    Microsoft case study, 572
    model assessment, 524, 572
    model construction, 524, 572
    neural network. *See* Neural network
    performance measures for binary classification, 576–582
    regression performance measures, 524–526
    regression trees, 538–548
  steps in, 524, 572
Data partitioning, 524, 526–535, 572, 573–576
Data preparation, 524, 526–535, 572, 573–576
Data query, 4
Data sampling, 524, 526–535, 572, 573–576
  class imbalanced data, 574–576
  holdout method, 526–529, 573–574
  k-fold cross-validation, 530–535, 574, 575
Data scientists, 8
Data security, 8
Data Tables (Excel), 640–642
Data Validation tool (Excel), 188
Data visualization, 83–150
  charts. *See* Charts
  Cincinnati Zoo & Botanical Gardens, 84–85
  data dashboards, 131–133
  data-ink ratio, 89–91
  geospatial data, 126–131
  heat maps, 120–121
  Movie industry case study, 150
  overview of, 86–91
  Pelican Stores case study, 149–150
  preattentive attributes, 86–89
  tables, 92–101
  treemaps, 121–122
Data wrangling, 151–198
  cleaning, 153, 167–176
  defined, 152
  discovery, 152–158
  enriching, 153, 176–186
  Procter & Gamble (P&G) case study, 152
  publishing, 153, 188
  structuring, 153, 158–166
  using Excel, 154–157

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Data wrangling (*Continued*)
Usman Solutions case study, 197–198
validating, 153, 186–188
Decision alternatives, 895
Decision analysis, 5, 893–942
branch probabilities, computation with Bayes' theorem, 910–913
with probabilities, 900–903
expected value approach, 900–902
risk analysis, 902–903
sensitivity analysis, 903
problem formulation, 895–897
decision trees, 896–897
payoff tables, 896
property purchase strategy case study, 939–940
with sample information, 904–910
expected value of perfect information, 909–910
expected value of sample information, 909
semiconductor fabrication at Axeon Labs, 941
U.S. Centers for Disease Control and Prevention example, 894
utility and, 914–918
utility theory, 913–922
without probabilities, 897–900
conservative approach, 898
minimax regret approach, 898–900
optimistic approach, 897–898
Decision boundary, 586, 587
Decision making
business analytics and, 894
data-driven, 2
defined, 3
managerial, 394
overview of, 3
uncertainty in, 672, 895
Decision nodes, 896
Decision strategy, 906–907
Decision trees, 896–897
Decision variables, 24, 638, 746–747, 764, 768
Decluttering, 90
Deep learning methods, 600
Degrees of freedom, 340
Delimiters, 154
Dendrograms, 277–280, 282
Dependent variable, 394, 420–421, 447, 494
Descriptive analytics, 4–5, 24
Descriptive data mining, 257–317
association rules, 284–287
automated recommender systems case study, 258
Big Ten Conference case study, 315–316
cluster analysis, 266–283
dimension reduction, 259–266
Know Thy Customer (KTC) case study, 316–317
text mining. *See* Text mining
Descriptive statistics, 21–82
African elephant populations case study, 81–82
data definitions and goals, 23–24
data distribution creation, 32–45
distribution analysis, 54–62
exploring data in excel, 27–32
Heavenly Chocolates case study, 80–81
measures of association between two variables, 62–68
measures of location, 46–51
measures of variability, 51–54
U.S. Census Bureau, 22
Diet problem, 779–781
Dimension reduction, 259–266
Discovery, 152–158
accessing data, 153–157
Discrete-event simulation, 716
Discrete probability distributions, 215–226, 736–742
custom, 215–217
expected value and variance, 217–220
risk analysis, 676
uniform, 220–221
Discrete random variables, 213–214, 217–220, 227
Discrete uniform distribution, 737
Distribution analysis, 54–62
boxplots, 59–62
empirical rule, 57–59
outlier identification, 59
percentiles, 55
quartiles, 56
z-score, 56–57, 58
Divisibility, linear optimization, 748
Documents, 288
computing dissimilarity between, 293–294
Double-subscripted decision variables, 769
Dow Jones Industrial Average, 23
Dow Jones Industrial Index, 23, 24–25
Dummy variables, 183–185, 424, 435
Duplicate records, 170–176

E

Efficient frontier, 865
eHarmony, 258
Embedded methods, 557–558, 610–611
Empirical probability distribution, 215
Empirical rule, 57–59
Enriching, 153, 176–186
data, 182–186
subsetting data, 177–179
supplementing data, 179–182
Ensemble methods, 543–548, 595–600
Entropy, 600
Equal-area cartogram, 131
Erroneous outliers, 170–176
Erroneous values, 170–176
Error Checking, 658, 659, 660
Estimated linear regression equation, 395–396
using Excel to compute, 401–402
Estimated multiple linear regression equation, 407
using Excel to compute, 409–412
Estimated regression line, 395
Ethical issues, in data and analytics usage, 12–14
Euclidean distance, 267–268
Evaluate Formulas, 658, 659
Events
chance, 895
complement of an event, 202–203
defined, 201
independent, 210
intersection of, 203
mutually exclusive, 204–205
and probabilities, 201–202
union of, 203
Evolutionary method
Excel's, heuristic optimization using, 869–877
Excel
AVERAGE function, 218, 357, 361
BINOM.DIST function, 223

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Excel (*Continued*)
charts. *See* Charts
chart tools, 401–402
coefficient of determination computation using, 406
CONCAT function, 39
COUNTA function, 347, 366
COUNTBLANK function, 168
COUNT function, 347, 357, 361
COUNTIF function, 42–43, 347, 366, 651–653
COUNTIFS function, 175
COVARIANCE.S function, 64, 65
data in
conditional formatting, 30–32
sorting and filtering, 27–30
Data Tables, 640–642
Data Validation tool, 188
data wrangling using, 154–157
estimated multiple linear regression equation using, 409–412
estimated simple linear regression equation using, 401–402
Evolutionary method, heuristic optimization using, 869–877
EXPON.DIST function, 238
exponential smoothing with, 492–494
forecasting with, 488–489
Forecast Sheet, 517–522
frequency distributions for quantitative data, 37
FREQUENCY function, 38–39, 43
frequency polygon, 42–43
GEOMEAN function, 50
Goal Seek, 642–644, 645
histograms, 37–41
HLOOKUP function, 180
hypothesis testing, 356–359, 365–366
IF function, 651–653
interval estimation using, 344–346
linear regression analysis using, 501, 503
MAX function, 52
MEDIAN function, 48
MIN function, 52
MODE.MULT function, 48
MODE.SNGL function, 48
NORM.DIST function, 234
NORM.INV function, 235–236
PERCENTILE.EXC function, 55
PivotCharts, 117–118
PivotTable, 97–101, 206–208
POISSON.DIST function, 225–226
RAND function, 678–681
random variables generation with, 677–681
Recommended Charts tool, 103–104
Regression tool, 416–417, 420, 421
Remove Duplicates tool, 174–176
simulation trials, 681, 682
spreadsheet modeling functions. *See* Spreadsheet models
STANDARDIZE function, 57
STDEV.S function, 54, 357, 361
SUM function, 650–651
SUMPRODUCT function, 217–218, 219, 650–651, 751–752
T.DIST function, 356
TEXTJOIN function, 166
variance calculation, 219–220
VLOOKUP function, 180
XLOOKUP function, 180, 654–656
Excel Solver
integer optimization problems, 810–814
linear programs, 751–754
nonlinear optimization problems, 853–855
overcoming local optima with, 858–859
Sensitivity Report, 762–764
Expected utility (EU), 917
Expected value (EV), 217–218, 900–902
of sample mean, 331
of sample proportion, 336
Expected value approach, 900–902
Expected value of perfect information (EVPI), 909–910
Expected value of sample information (EVSI), 909
Experimental region, 399
Experiments, random, 201
EXPON.DIST function, 238
Exponential distribution, 735
Exponential probability distribution, 236–239
Exponential smoothing, 485, 490–494
Exponential utility function, 921–922
Extrapolation, 399
Extreme points, 750

F

Facebook, 258
False negative, 577
False positive, 577
Feasibility table, 826
Feasible regions
extreme points of, 750
integer linear optimization models, 809
linear optimization models, 749
nonlinear optimization models, 852–853
Feasible solution, 749
Features, 572
Feature (variable) selection, 259, 555–558
defined, 609
embedded methods, 557–558, 610–611
filter methods, 556–557, 610
wrapper methods, 556, 609
Feed forward neural network, 548–549, 601
Field, 157
Filter methods, 556–557, 610
Financial analytics, 10
Finite population, sampling from, 322–323
Finite population correction factor, 332
Fitted distribution, bid fraction data, 700–708
Fixed cost, 635
problem, 816–820
Flat file, 157
Forecast error, 482–483
Forecasting, 472–473
ACCO Brands Corporation case study, 472
accuracy, 481–485
exponential smoothing forecasting, 492–494
moving averages forecasting, 489–490
Bass forecasting model, 866–869
causal/exploratory methods, 473
Excel Forecast Sheet, 517–522
exponential smoothing, 485, 490–494
food and beverage sales case study, 515
linear regression analysis for, 494–504
causal forecasting, 500–503
combining causal variables with trend and seasonality effects, 503–504
linear trend projection, 494–496
seasonality without trend, 496–497
seasonality with trend, 497–500

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Forecasting (*Continued*)
lost sales case study, 515–516
models, determination for usage, 504–505
moving averages method, 485, 486–490
qualitative methods, 473
quantitative methods, 473
Forecast Sheet (Excel), 517–522
Forward selection procedure, 441, 442
Fowle Marketing Research problem, 776–779
Frame, 321
Frequency distributions
for categorical data, 32–33
cumulative, 43–45
percent, 33–35
for quantitative data, 35–37
relative, 33–35
Frequency document-term matrix, 290
FREQUENCY function, 38–39, 43
Frequency polygons, 42–43
F1 Score, 581

G

Galton, Francis, 396
Gamma distribution, 735
Gebhardt Electronics case study, 256
General Data Protection Regulation (GDPR), 13
GEOMEAN function, 50
Geometric approach
integer linear optimization models, 808–810
to linear program solution, 749–751
Geometric mean, 48–51
Geospatial data
cartograms, 129–131
choropleth maps, 127–129
defined, 126
visualization, 126–131
Gini index, 600
Global maximum, 856
Global minimum, 856
Global optimum, nonlinear optimization problems, 856–858
Goal Seek (Excel), 642–644
Government, use of analytics by, 11–12
Greater-than-or-equal-to constraint, 763
Grey Code Corporation (GCC) case study, 630–632
Group average linkage, 276
Growth factor, 49

H

Hadoop, 8
Half spaces, 749
Hamilton County judges case study, 254–255
Health care analytics, 11
Heat maps, 120–121
Heavenly Chocolates case study, 80–81
Heuristic method, 870
Heuristic optimization using Excel's Evolutionary method, 869–877
Hidden layer, neural network, 548, 601
Hierarchical clustering, 266, 275–283
k-means clustering *vs.,* 283
and measuring dissimilarity between clusters, 275–282
Hierarchical data, 121
High-low-close stock chart, 125
Histograms, 37–41
HLOOKUP function, 180
Holdout method, 443
Horizontal pattern, time series analysis, 474–476
Housing bubble case study, 568–569
Human resource (HR) analytics, 10
Hypergeometric distribution, 738–739, 740
Hypothesis tests, 349–367, 413
alternative hypotheses, 349–352
big data and, 374–376
interval estimation and, 363–364
null hypotheses, 349–352
one-tailed tests, 353–354
of population mean, 351–352
in population proportion, 351–352
of population proportion, 364–367
steps of, 362
summary and practical advice, 362
test statistic, 354–359
two-tailed test, 359–362
Type I and Type II errors, 352–353
using Excel, 356–359

I

Illegitimately missing data, 167
Impurity, 591
Imputation, 168
Independent events, 210
multiplication law for, 210
Independent variables, 394, 446, 494
categorical, 424–428
interaction between, 436–440
nonsignificant, 420–421
variable selection procedures, 441–442
Infeasibility, 759–760
Inference. *See* Statistical inference
Infinite population, sampling from, 323–325
Influence diagrams, 635–637
Input layer, neural network, 548, 601
Institute for Operations Research and the Management Sciences (INFORMS), 13, 14
Integer linear optimization models, 805–847. *See also* Linear optimization models
binary variables. *See* Binary variables
Eastborne Realty example, 807–814
Excel Solver, 810–814
geometry of, 808–810
sensitivity analysis, 813–814
types of, 806–807
Vattenfall example, 806
Integer linear programs, 806
Integer optimality, 814
Integer uniform distribution, 737
Interaction, between independent variables, 436–440
Interquartile range (IQR), 56
Intersection, of half spaces, 749
Intersection of events, 203
Interval estimation, 339–349, 413
defined, 339
and hypothesis tests, 363–364
of population mean, 339–346
of population proportion, 346–348
using Excel, 344–346

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Inventory policy, analysis for Promus Corp, 686–693
Investment portfolio selection, 765–768
Investment strategy case study, 801–803
Itemset, 284

J

Jaccard distance, 269–270
Jaccard's coefficient, 269
John Morrell & Company case study, 320
Joint probabilities, 208

K

Key performance indicators (KPIs), 131–132
k-fold cross-validation, 530–535, 574, 575
k-means clustering, 266, 271–275
  hierarchical clustering *vs.,* 283
k-nearest neighbor (k-NN) classification, 587–590
k-nearest neighbors (k-NN) regression, 535–538
Knot, nonlinear relationships, 434
Know Thy Customer (KTC) case study, 316–317
k out of *n* alternatives constraint, 826
Kroger, 8

L

Lagged variable, 503
Lagrangian multiplier, 855
Land Shark Inc. example, 693–708
  execution of simulation trials and output analysis, 698–700
  generating bid amounts with fitted distributions, 700–708
  generating values for random variables, 696–698
  spreadsheet model for, 694–696
Lasso regularization, 557, 611
Law of Large Numbers, 445
Lazy learner, 535, 587
Leaf node, 541, 593
Least squares linear regression model, 413–417
Least squares method, 397–402
  equation, 398
  estimates of regression parameters, 399–401
  and multiple linear regression, 408
Leave-one-out cross-validation, 530, 574
Legal issues, in data and analytics usage, 12–14
Legitimately missing data, 167
Less-than-or-equal-to constraints, 778
Level of significance, 343, 353, 359, 363
Lift chart, 579–580, 581
Lift ratio, 285
Linear functions, 748
Linear optimization models, 743–805. *See also* Integer linear optimization models
  alternative optimal solutions, 758–759, 781–782
  applications of, 744
  assignment problem, 776–779
  constraints, 744, 746–748
  decision variables, 746–747
  diet problem, 779–781
  infeasibility, 759–760
  investment portfolio selection, 765–768
  investment strategy case study, 801–803
  linear programming notation and examples, 764–781
  linear programming outcomes, 757–762
  maximization problem. *See* Maximization problem
  maximizing banner ad revenue, 772–776
  M&D Chemicals problem, 755–757, 765
  minimization problem, 755–757
  Par, Inc. problem, solving, 749–754
  sensitivity analysis, 762–764
  Stitch Fix example, 744
  transportation problem, 768–772
  unbounded solutions, 760–761
Linear programs/programming, 745, 748
  Excel Solver for, 751–754
  geometric approach to solving, 749–751
  notation and examples, 764–781
Linear regression, 393–470
  alumni giving case study, 466–467
  categorical independent variables, 424–428
  Consumer Research, Inc. case study, 468
  inference and, 412–424
    least squares linear regression model, 413–417
    multicollinearity, 421–423
    nonsignificant independent variables, 420–421
    very large samples, 443–446
  model fitting, 441–443
    overfitting, 442–443
    variable selection procedures, 441–442
  modeling nonlinear relationships, 429–441
    interaction between independent variables, 436–440
    piecewise linear regression models, 434–436
    quadratic regression models, 430–434
  multiple. *See* Multiple linear regression
  NASCAR case study, 469–470
  prediction with, 447–450
  simple. *See* Simple linear regression model
  Walmart.com, 394
Linear regression analysis, for forecasting, 494–504
  causal forecasting, 500–503
  combining causal variables with trend and seasonality effects, 503–504
  linear trend projection, 494–496
  seasonality without trend, 496–497
  seasonality with trend, 497–500
Linear trend projection, 494–496
Line charts, 92, 93, 104–108
LinkedIn, 258
Local maximum, 856
Local minimum, 856
Local optimum, nonlinear optimization problems, 856–859
Location problem
  integer linear optimization models, 820–822, 827–829
  nonlinear optimization model, 860–861
Logistic function, 585, 605
Logistic regression, classification with, 582–587
Logistic S-curve, 583, 584
Log loss, 606
Log-normal distribution, 736
Lower-tail test, 359
LP relaxation, 807

M

MagicBand, 9
Maillard Riposte, protein consumption for, 261–266
Make-versus-buy decision, 635, 638
Managerial decisions, 394
Manhattan distance, 268–269

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Manufacturer's product information, 351
MapReduce, 8
Maps
  cartogram, 129–131
  choropleth, 127–129
  heat maps, 120–121
  treemaps, 121–122
Marginal probabilities, 208
Margin of error, 339, 344
Market basket analysis, 284
Marketing analytics, 10–11
Market segmentation, 266
Markowitz mean-variance portfolio model, 861–865
Match.com, 258
Matching coefficient, 269
Matching distance, 269–271
Mathematical models, 635–637, 748
MAX function, 52
Maximization problem, 745–748
  mathematical model, 748
  problem formulation/modeling, 746–748
Maximizing banner ad revenue, 772–776
Maximum Time without Improvement parameter, Evolutionary method, 871
McNeil's Auto Mall case study, 255–256
McQuitty's linkage, 283
Mean
  arithmetic, 46, 47
  deviation about the, 52
  geometric, 48–51
  population. *See* Population mean
  sample. *See* Sample mean ($\bar{x}$)
Mean absolute error (MAE), 483, 485, 489, 525
Mean absolute percentage error (MAPE), 483–484, 485, 489
Mean forecast error (MFE), 482–483
Mean squared error (MSE), 483, 485, 489–490, 525–526
Measurement error, 369
Measures of association, intervariable, 62–68
  correlation coefficient, 67–68
  covariance, 64–66
  scatter charts, 62–63, 66
Measures of location, 46–51
  geometric mean, 48–51
  mean (arithmetic mean), 46, 47
  median, 47–48
  mode, 47, 48
Measures of variability, 51–54
coefficient of variation, 54
range, 51–52
standard deviation, 53–54
variance, 52–53
Median, 47–48
MEDIAN function, 48
Median linkage, 283
Microsoft case study, 572
MIN function, 52
Minimax regret approach, 898–900
Minimization problem, 755–757
Min–max scaling, 551, 552, 603–604
Missing at random (MAR), 167, 168
Missing completely at random (MCAR), 167, 168, 169
Missing data, 167–170
Missing not at random (MNAR), 167, 168
Mixed-integer linear program, 807
Mode, 47, 48
Model assessment, 572
Model construction, 572
Modeling, 746–748
MODE.MULT function, 48
MODE.SNGL function, 48
Monte Carlo simulation, 671–742
  advantages and disadvantages of, 714–715
  analysis, steps for conducting, 715–716
  defined, 672
  with dependent random variables, 709–714
  financial risk for loan providers, evaluation of, 672
  fitted distribution, 700–708
  inventory policy analysis for Promus Corp, 686–693
  Land Shark Inc. example, 693–708
  output analysis, 691–693, 698–700
  random variables. *See* Random variables
  risk analysis. *See* Risk analysis
  Sanotronics LLC example, 673–686
  simulation modeling, 693–708
  spreadsheet model, 687–688, 694–696
  validation, 714
  verification, 714
Movie industry case study, 150
Moving averages method, 485, 486–490
Multicollinearity, 421–423
Multimodal data, 48
Multiple-choice constraint, 825–826
Multiple linear regression, 394
  Butler Trucking Company and, 408–409
  estimated multiple linear regression equation, 407
  estimation process for, 408
  least squares method and, 408
  model, 407–412
  using Excel, 409–412
Multiplication law, 210
Music Genome Project, 258
Mutation Rate parameter, Evolutionary method, 871
Mutually exclusive constraint, 825–826
Mutually exclusive events, 204–205

N

Naïve forecasting method, 481
NASCAR case study, 469–470
National Aeronautics and Space Administration (NASA), 200
Negative binomial distribution, 739, 741
Netflix, 258
Networks, 768
Neural network
  defined, 548, 600
  feed forward, 601
  overview, 600–601
  perceptron, 601–602
  regression, 548–555
  structure of, 548–552, 601–605
  training process, 552–555, 605–608
Neurons, 548, 601
Neuron weight, 549, 550, 554, 602, 603, 607, 608
New product adoption, Bass forecasting model, 866–869
Nodes, 635, 768, 896
Nonlinear optimization model, 849–892
  AESA example, 850
  Bass forecasting model, 866–869
  global optimum, 856–858

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Nonlinear optimization model (*Continued*)
heuristic optimization using Excel's Evolutionary method, 869–877
local optimum, 856–859
location problem, 860–861
Markowitz mean-variance portfolio model, 861–865
portfolio optimization with transaction costs case study, 889–892
production application, 850–856
constrained problem, 851–853
Excel Solver, 853–855
sensitivity analysis, 855–856
shadow prices, 855–856
unconstrained problem, 850–851
Nonlinear optimization problem, 850
Nonlinear relationships
modeling, 429–441
interaction between independent variables, 436–440
piecewise linear regression models, 434–436
quadratic regression models, 430–434
Nonnegativity constraints, 748, 749
Nonprofit organizations, use of analytics by, 11–12
Nonresponse error, 368–369
Nonsampling error, 368–369
Normal distribution, 677, 734
Normal probability density function, 232
Normal probability distributions, 231–236
NORM.DIST function, 234
NORM.INV function, 235–236
Null hypotheses, 349–352
Numeric data, 159

O

Objective function, 744, 764
Objective function contour, 750
Observations, 23, 258
measuring distance between, 267–271
Observed level of significance, 359
Ockham's razor, 443
Odds, 583
OKCupid, 258
One-tailed tests, 353–354
One-to-many merger, 179
One-to-one merger, 179
One-way data table, 640, 641
Operational decisions, 3
Opportunity loss, 898, 899
Optimistic approach, 897–898
Optimization models, 5
applications of, 744
integer linear. *See* Integer linear optimization models
linear. *See* Linear optimization models
nonlinear. *See* Nonlinear optimization model
Orbitz case study, 524
Order-up-to point, 652
Order-up-to policy, 652
Outcomes, 201–202, 895
Outliers
in boxplots, 60–61
identifying, 59
Out-of-bag estimation, 544, 546, 547, 596, 598–599
Output layer, neural network, 548, 601
Overall error rate, 577
Overfitting, 442–443, 526, 573
Oversampling approach, 576

P

Pandora Internet Radio, 258
Par, Inc. problem
Excel Solver for, 751–754
feasible region for, 749
feasible solution for, 749
geometry of, 749–751
nonlinear optimization model, 850–856
solving, 749–754
Parallel-coordinates chart, 126
Parameters, 322, 638, 639
population, 395
Part-worths, 822–824
Payoffs, 896
Payoff tables, 896
Pearson, Karl, 396
Pelican Stores case study, 149–150
Percent frequency distributions, 33–35
PERCENTILE.EXC function, 55
Percentiles, 55
Perceptron, 549, 601–602
Perfect information, 909–910
Piecewise linear regression models, 434–436
Pie charts, 112
PivotCharts (Excel), 117–118
PivotTables (Excel), 97–101, 206–208
Point estimation, 326–328
Point estimator, 327, 395
unbiased, 331
POISSON.DIST function, 225–226
Poisson distribution, 739, 742
Poisson probability distribution, 224–226
Poisson probability mass function, 224
Population, 24
characteristics of, 320
finite, 322–323
infinite, 323–325
with normal distribution, 333
point estimator of, 326–328
sampled, 321, 328
target, 328
without normal distribution, 333
Population mean
hypothesis tests in, 351–352, 353–364
interval estimation of, 339–346
Population parameters, 395
Population proportion
hypothesis tests in, 351–352
hypothesis tests of, 365–366
interval estimation of, 346–348
Population Size parameter, Evolutionary method, 871
Posterior probabilities, 211, 904, 905, 912
Postoptimality analysis, 762
Practical significance, 374
of statistical inference, 367–376
Preattentive attributes, 86–89
Precision, 581
Prediction error, 525
Prediction interval, 448–450
Predictive analytics, 5
Predictive data mining, 523–632. *See also* Data mining
classification trees, 591–600
classification with logistic regression, 582–587
data partitioning, 524, 526–535, 572, 573–576

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Predictive data mining (*Continued*)
data preparation, 524, 526–535, 572, 573–576
data sampling, 524, 526–535, 572, 573–576
feature selection, 555–558, 609–611
housing bubble case study, 568–569
k-nearest neighbor classification, 587–590
k-nearest neighbor regression, 535–538
Microsoft case study, 572
model assessment, 524, 572
model construction, 524, 572
neural network. *See* Neural network
performance measures for binary classification, 576–582
regression performance measures, 524–526
regression trees, 538–548
Predictive spreadsheet model, 660–661
Prescriptive analytics, 5–6, 744
Prescriptive spreadsheet model, 660–661
Presence/absence document-term matrix, 288–289
Press Teag Worldwide (PTW) example, 709–714
Principal component analysis, 259
geometric interpretation of, 259–262
Prior probabilities, 211, 904, 905
Probabilities, 199–256
addition law, 203–205
basic relationships of, 202–205
branch, computation with Bayes' theorem, 910–913
case studies, 254–256
classification, 578
conditional, 205–213, 912
continuous probability distributions, 227–239
cumulative, 231
defined, 200
discrete probability distributions, 215–226
events and, 201–202
joint, 208
marginal, 208
National Aeronautics and Space Administration, 200
posterior, 211, 904, 905, 912
prior, 211, 904, 905
random variables, 213–215
Probability density function, 227
Probability distributions, 673
binomial, 221–223
continuous. *See* Continuous probability distributions
defined, 215
discrete. *See* Discrete probability distributions
empirical, 215
exponential, 236–239
normal, 231–236, 677, 734
Poisson, 224–226
for random variables, 734–742
to represent random variables, 676–677
triangular, 229–231
uniform, 676, 677, 701
Probability mass function, 215
Problem formulation, 746–748, 755, 895–897
decision trees, 896–897
payoff tables, 896
Procter & Gamble (P&G) case study, 152, 634
Product design and market share optimization problem, 822–825
Promus Corp example
demand, value generation for, 688–690
inventory policy analysis for, 686–693
spreadsheet model for, 687–688
Proportionality, linear optimization, 748
Pruning, 542–543, 594, 595
*p*th percentile, 55
Publishing, 153, 188
p values, 418–419
big data and, 374–375
hypothesis tests, 355–362
nonlinear relationships, 434
very large samples, 444–446
Pythagorean theorem, 260–261

Q

Quadratic function, 851
Quadratic regression models, 430–434
Qualitative forecasting methods, 473
Quality Associates, Inc. case study, 391–392
Quantitative data, 24
frequency distributions for, 35–37
frequency polygons, 42–43
histograms, 37–41
Quantitative forecasting methods, 473
Quartiles, 56
Quick Analysis button (Excel), 32

R

RAND function, 678–681
Random experiments, 201
Random forest, 547, 599–600
Random sampling, 24, 322–325
Random Seed parameter, Evolutionary method, 871
Random variables, 24, 213–215, 328
beta distribution, 706–708
continuous, 214–215, 227
dependent, 709–714
discrete, 213–214, 217–220, 227
expected value and variance, 217–218
generating values for Land Shark, 696–698
generation, with Excel, 677–681
Monte Carlo simulation, 673, 676–677
probability distributions for, 734–742
probability distributions to represent, 676–677
Range, 51–52
Raw data, 152
format of, 157–158
Recall, 581
Receiver operating characteristic (ROC) curve, 581–582
Recommended Charts tool (Excel), 103–104
Records, 157, 258
duplicate, 170–176
Rectified linear unit (ReLu) function, 551–552, 604–605
Reduced cost, 764
Reduced gradient, 855
Regression
defined, 394, 524
k-nearest neighbors, 535–538
neural network, 548–555
performance measures, 524–526
Regression analysis
linear, for forecasting, 494–504
causal forecasting, 500–503
combining causal variables with trend and seasonality effects, 503–504
linear trend projection, 494–496
seasonality without trend, 496–497
seasonality with trend, 497–500

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Regression lines, 395–396
Regression parameters, estimates of, 399–401
Regression trees, 538–548
  constructing, 538–541
  defined, 538
  ensemble methods, 543–548
  generating predictions with, 541–543
Regret, 898, 899
Regularization, 557–558, 610
Rejection rule, 358
Relative frequency distributions, 33–35
Remove Duplicates tool (Excel), 174–176
Research hypothesis, 349–350
Residual, 398
Residual plots, logistic regression, 584
Response variable. *See* Dependent variable
Ridge regularization, 557, 611
Right-hand side allowable increase (decrease), 764
Risk analysis
  base-case scenario, 673–674
  best-case scenario, 674
  in decision analysis, 902–903
  defined, 673
  generating values for random variables with Excel, 677–681
  Monte Carlo simulation, 673–686
  probability distributions to represent random variables, 676–677
  simulation output, measurement and analysis, 682–686
  simulation trials with Excel, 681, 682
  spreadsheet models, 674–675
  worst-case scenario, 674
Risk avoider, 915
Risk neutral, 921
Risk profile, 902–903
Risk taker, 918
Root mean squared error (RMSE), 525–535
Root node, 539, 591
Rule-based models, 5

S

Sales forecasting. *See* Forecasting
Sampled population, 321, 328
Sample information, 904
  decision analysis with, 904–910
  expected value of, 909
Sample mean ($\bar{x}$), 26, 321, 328–331
  expected value of, 331
  sampling distribution of, 331–336, 339–340
  standard deviation of, 331–332
Sample proportion, 328–331
  expected value of, 336
  sampling distribution of, 336–339
  standard deviation of, 336–337
Sample/sampling, 24, 321
  distributions, 328–339
    of sample mean, 331–336, 339–340
    of sample proportion, 336–339
    sample size and, 334–336
  from finite population, 322–323
  from infinite population, 323–325
  random, 24, 322–325
  selection of, 322–326
  very large samples, 443–446
Sample size, 334–336, 338–339
Sample space, 201
Sample statistic, 326
Sampling error, 332, 367–368
  big data and, 371–372
Scatter-chart matrix, 115–116
Scatter charts, 62–63, 66, 101–103
logistic regression, 583
Scenario Manager (Excel), 644–649
Scree plot, 263
Seasonality
    effects, combining causal variables with trend and, 503–504
    with trend, 497–500
    without trend, 496–497
Seasonal pattern, time series analysis, 477–480
Semi-structured data, 158
Sensitivity, 581
Sensitivity analysis
    cautionary note about, 813–814
    in decision analysis, 903
    defined, 762
    Excel Solver Sensitivity Report, interpretation of, 762–764
    nonlinear optimization problems, 855–856
Sensitivity Report (Excel Solver), 762–764
Shadow prices, 763, 855–856
Show Formulas, 656
Significance tests, 353
Simple linear regression model, 394, 395–396
  estimated equation, 395–396
  fit assessment, 403–406
  least squares method, 397–402
Simple random sample, 322–323
Simulation modeling, 693–708
Simulation optimization, 5
Simulations
  defined, 5
  Monte Carlo. *See* Monte Carlo simulation
Simulation trials
  with Excel, 681, 682
  execution, 691–693, 698–700
Single linkage, 275–276
Skewness, 41
Slack value, 753–754
Slack variable, 754
Smoothing constant, 490
Smoothing methods
  exponential smoothing, 485, 490–494
  moving averages, 485, 486–490
Sparklines, 106–108
Specificity, 581
Sports analytics, 12
Spreadsheet models, 633–670
  auditing. *See* Auditing, spreadsheet models
  building, 635–640
  Excel functions for, 649–656
    COUNTIF function, 651–653
    IF function, 651–653
    SUM function, 650–651
    SUMPRODUCT function, 650–651
    XLOOKUP function, 654–656
  formatting, 639
  influence diagrams, 635–637
  for Land Shark Inc., 694–696
  make-versus-buy decision, 635, 638
  mathematical models, 635–637
  overview, 634
  parameters, 638, 639

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Spreadsheet models (*Continued*)
predictive, 660–661
prescriptive, 660–661
for Press Teag Worldwide, 709–713
Procter & Gamble case study, 634
for Promus Corp, 687–688
retirement plan case study, 670
risk analysis, 674–675
what-if analysis. *See* What-if analysis
Stacked-column chart, 110, 111
Stacked data, 161
Standard deviation, 53–54
of sample mean, 331–332
of sample proportion, 336–337
Standard error, 331, 337
Standard error of mean, 337
Standard error of proportion, 337
Standardized value, 57
STANDARDIZE function, 57
Standard normal distribution, 340, 341
States of nature, 895–896
Static holdout method, 526–529, 573–574
Stationary time series, 474
Statistical inference, 319–392
big data and. *See* Big data
defined, 321, 412
hypothesis tests. *See* Hypothesis tests
interval estimation. *See* Interval estimation
John Morrell & Company case study, 320
and linear regression, 412–424
least squares linear regression model, 413–417
multicollinearity, 421–423
nonsignificant independent variables, 420–421
very large samples, 443–446
nonsampling error, 368–369
point estimation, 326–328
practical advice for, 328
practical significance of, 367–376
Quality Associates, Inc. case study, 391–392
sample selection, 322–326
sampling distributions, 328–339
sampling error, 367–368
Young Professional magazine case study, 390–391
STDEV.S function, 54, 357, 361
Stemming, 290
Stepwise selection procedure, 441, 442
Stitch Fix case study, 744
Stock charts, 124–126
Stopwords, 290
Strategic decisions, 3
Stratified random sampling, 575
Structured data, 157
Structuring, 153, 158–166
arrangement of data, 159–161
combining multiple fields into a single field, 165–166
data formatting, 159
splitting a single field into multiple fields, 161–165
Stump, 544, 595
Subsetting data, 177–179
SUM function, 650–651
Sum of squares, 403–405
Sum of squares due to error (SSE), 403
Sum of squares due to regression (SSR), 404–405
SUMPRODUCT function, 217–218, 219, 650–651, 751–752
Supervised learning, 5, 524, 572. *See also* Predictive data mining
Supplementing data, 179–182
Supply chain analytics, 11
Support, 284
Surplus variable, 757

T

Table lens, 116–117
Tables, 92–101
charts *vs.,* 92–93
crosstabulation, 94–96
design principles, 93–94
PivotTables (Excel), 97–101, 206–208
Tactical decisions, 3
Tall data, big data as, 370
Target population, 328
T.DIST function, 356
t distributions, 340–341, 355
Term frequency times inverse document frequency (TFIDF), 291–293
Term normalization, 290
Terms, 288
Test of independence, 610
Test set, 528, 529
Test statistic, 354–359
Text data, 159
TEXTJOIN function, 166
Text mining, 287–295
computing dissimilarity between documents, 293–294
defined, 287
movie reviews, 290–293
preprocessing text data for analysis, 289–290
unstructured data, 287
voice of the customer at Triad Airline, 288–289
word clouds, 294–295
Three-dimensional charts, 112
Time series, defined, 474
Time series analysis
forecasting and, 473
patterns, 474–481
cyclical, 481
horizontal, 474–476
identification of, 481
seasonal, 477–480
trend, 476–477, 478–480
Time series data, 24–25
Time series plot, 474, 475
Tokenization, 289–290
Total sum of squares (SST), 403–404
Trace Dependents (Excel), 656, 657
Trace Precedents (Excel), 656, 657
Training set, 443, 526, 527, 536, 554, 573–574, 588
Transportation problem, 768–772
Treemaps, 121–122
Trend-cycle effects, 481
Trendline, 102
Trend pattern, time series analysis, 476–477, 478–480
Triad Airlines example, 288–289
Trial-and-error approach, 639
Trials, 677
Triangular distribution, 701–705, 735–736
Triangular probability density function, 230
Triangular probability distributions, 229–231
t test, 417–418
Two-tailed test, 359–362

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Two-way data table, 640, 643
Type I error, 352–353, 359
Type II error, 352–353

U

Unbiased point estimator, 331
Unbounded solutions, in linear programming problems, 760–761
Uncertainty, 200, 672, 895
Uncertain variables, 24
Unconstrained problem, nonlinear optimization models, 850–851
Undersampling approach, 576
Uniform probability density function, 228
Uniform probability distributions, 220–221, 227–229, 676, 677, 701, 736
Union of events, 203, 204
Unstable learner, 543, 595–596
Unstacked data, 161
Unstructured data, 157, 287
Unsupervised learning, 4
Unsupervised learning techniques, 258. *See also* Descriptive data mining
Upper-tail test, 359
U.S. Census Bureau, 22
Usman Solutions case study, 197–198
Utility
  and decision analysis, 914–918
  defined, 913
Utility functions, 918–921
  exponential, 921–922
Utility theory, 5, 913–922
  exponential utility function, 921–922
  utility and decision analysis, 914–918
  utility functions, 918–921

V

Validating, 153, 186–188
Validation, 714
  leave-one-out cross-validation, 530
Validation set, 443, 526, 527, 528, 536, 554
Variable cost, 635
Variables
  binary. *See* Binary variables
  categorical, 185–186, 496–497
  causal, 503–504
  decision, 24, 638, 746–747, 764, 768
  defined, 23
  dependent/response, 394, 420–421, 447, 494
  dummy, 183–185, 424, 435
  independent/predictor. *See* Independent variables
  lagged, 503
  measures of association between two, 62–68
  random. *See* Random variables
  selection procedures, 441–442
  slack, 754
  surplus, 757
  uncertain, 24
Variance, 52–53, 219–220
Variation, 24
Variety, big data, 7, 370
Vattenfall case study, 806
Velocity, big data, 6–7, 370
Venn diagrams, 202–203
Veracity, big data, 7–9, 370
Verification, 714
Virtual item, 287
VLOOKUP function, 180
Volume, big data, 6, 7, 370

W

Walmart.com, 394
Walt Disney Company, 9
Ward's linkage, 276–277
Watch Window (Excel), 659, 660
Waterfall charts, 122–124
Weak learner, 547, 600
Web analytics, 12
What-if analysis, 634, 640–649
  Data Tables (Excel), 640–642
  Excel Solver, 751–754
  Goal Seek (Excel), 642–644, 645
  risk analysis, 674
  Scenario Manager (Excel), 644–649
Wide data, big data as, 370
Word clouds, 294–295
Worst-case scenario, 674
Wrapper methods, 556, 609

X

XLOOKUP function, 180, 654–656

Y

y-intercept, 399, 420
Young Professional magazine case study, 390–391
YouTube, 258

Z

z-scores, 56–57, 58, 268

Copyright 2024 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.