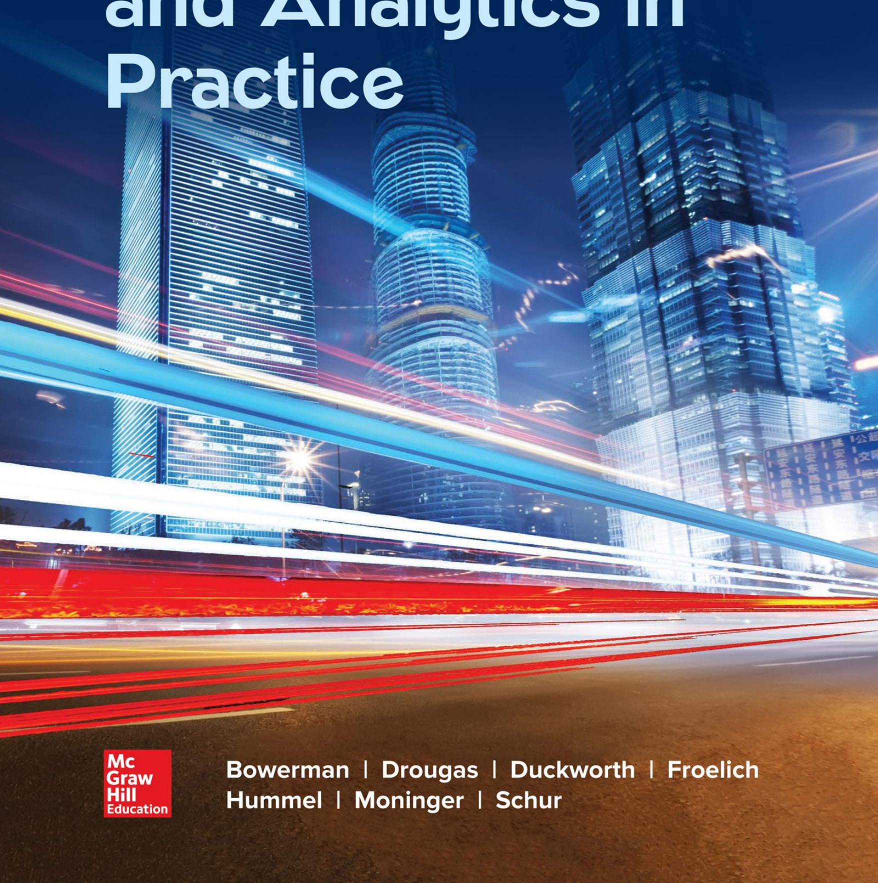
Ninth Edition

Business Statistics and Analytics in Practice



Mc Graw Hill Education

Bowerman | Drougas | Duckworth | Froelich
Hummel | Moninger | Schur

Bruce L. Bowerman
Miami University

Anne M. Drougas
Dominican University

William M. Duckworth
Creighton University

Amy G. Froelich
Iowa State University

Ruth M. Hummel
JMP

Kyle B. Moninger
Bowling Green State University

Patrick J. Schur
Miami University

Business Statistics and Analytics in Practice

NINTH EDITION

with major contributions by

Steven C. Huchendorf
University of Minnesota

Dawn C. Porter
University of Southern California

Mc
Graw
Hill
Education

BUSINESS STATISTICS AND ANALYTICS IN PRACTICE, NINTH EDITION

Published by McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 LWI/LWI 21 20 19 18

ISBN 978-1-260-18749-6
MHID 1-260-18749-7

Portfolio Manager: *Noelle Bathurst*
Product Developer: *Tobi Philips*
Marketing Manager: *Harper Christopher*
Content Project Managers: *Fran Simon/Jamie Koch*
Buyer: *Laura Fuller*
Design: *Egzon Shaqiri*
Content Licensing Specialists: *Brianna Kirschbaum*
Cover Image: ©*hxdyl/Shutterstock*
Compositor: *MPS Limited*

All credits appearing on page or at the end of the book are considered to be an extension of the copyright page.

Library of Congress Cataloging-in-Publication Data

Bowerman, Bruce L., author. | Duckworth, William Maury, author. |
Froelich, Amy Goodwin, author.
Business statistics in practice / Bruce L. Bowerman, Miami University,
William M. Duckworth, Creighton University, Amy Froelich, Iowa State University.
Ninth Edition. | Dubuque : McGraw-Hill Education, [2019] |
Revised edition of Business statistics in practice, [2017]
LCCN 2018012294 | ISBN 9781260187496 (alk. paper)
LCSH: Commercial statistics. | Statistics.
LCC HF1017 .B654 2018 | DDC 519.5024/65—dc23
LC record available at https://lccn.loc.gov/2018012294

The Internet addresses listed in the text were accurate at the time of publication. The inclusion of a website does not indicate an endorsement by the authors or McGraw-Hill Education, and McGraw-Hill Education does not guarantee the accuracy of the information presented at these sites.

mheducation.com/highered

The McGraw-Hill/Irwin Series in Operations and Decision Sciences

SUPPLY CHAIN MANAGEMENT

Benton
Purchasing and Supply Chain Management
Second Edition

Burt, Petcavage, and Pinkerton
Supply Management
Eighth Edition

Bowersox, Closs, Cooper, and Bowersox
Supply Chain Logistics Management
Fourth Edition

Johnson and Flynn
Purchasing and Supply Management
Fifteenth Edition

Simchi-Levi, Kaminsky, and Simchi-Levi
Designing and Managing the Supply Chain: Concepts, Strategies, Case Studies
Third Edition

PROJECT MANAGEMENT

Brown and Hyer
Managing Projects: A Team-Based Approach
First Edition

Larson and Gray
Project Management: The Managerial Process
Seventh Edition

SERVICE OPERATIONS MANAGEMENT

Bordoloi, Fitzsimmons, and Fitzsimmons
Service Management: Operations, Strategy, Information Technology
Ninth Edition

MANAGEMENT SCIENCE

Hillier and Hillier
Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets
Sixth Edition

BUSINESS RESEARCH METHODS

Schindler
Business Research Methods
Thirteenth Edition

BUSINESS FORECASTING

Keating and Wilson
Forecasting and Predictive Analytics with ForecastX^{TM}
Seventh Edition

LINEAR STATISTICS AND REGRESSION

Kutner, Nachtsheim, and Neter
Applied Linear Regression Models
Fourth Edition

BUSINESS SYSTEMS DYNAMICS

Sterman
Business Dynamics: Systems Thinking and Modeling for a Complex World
First Edition

OPERATIONS MANAGEMENT

Cachon and Terwiesch
Operations Management
First Edition

Cachon and Terwiesch
Matching Supply with Demand: An Introduction to Operations Management
Fourth Edition

Jacobs and Chase
Operations and Supply Chain Management
Fifteenth Edition

Jacobs and Chase
Operations and Supply Chain Management: The Core
Fourth Edition

Schroeder and Goldstein
Operations Management in the Supply Chain: Decisions and Cases
Seventh Edition

Stevenson
Operations Management
Thirteenth Edition

Swink, Melnyk, Hartley, and Cooper
Managing Operations Across the Supply Chain
Third Edition

BUSINESS MATH

Slater and Wittry
Practical Business Math Procedures
Twelfth Edition

Slater and Wittry
Math for Business and Finance: An Algebraic Approach
Second Edition

BUSINESS STATISTICS

Bowerman, et. al.
Business Statistics and Analytics in Practice
Ninth Edition

Bowerman, O'Connell, Murphree, and Orris
Essentials of Business Statistics
Fifth Edition

Doane and Seward
Applied Statistics in Business and Economics
Sixth Edition

Doane and Seward
Essential Statistics in Business and Economics
Second Edition

Jaggia and Kelly
Business Statistics: Communicating with Numbers
Third Edition

Jaggia and Kelly
Essentials of Business Statistics: Communicating with Numbers
First Edition

Lind, Marchal, and Wathen
Basic Statistics for Business and Economics
Ninth Edition

Lind, Marchal, and Wathen
Statistical Techniques in Business and Economics
Seventeenth Edition

McGuckian
Connect Master: Business Statistics

ABOUT THE AUTHORS

Courtesy of Bruce Bowerman

Bruce L. Bowerman Bruce L. Bowerman is emeritus professor of information systems and analytics at Miami University in Oxford, Ohio. He received his Ph.D. degree in statistics from Iowa State University in 1974, and he has over 40 years of experience teaching basic statistics, regression analysis, time series forecasting, survey sampling, and design of experiments to both undergraduate and graduate students. In 1987 Professor Bowerman received an Outstanding Teaching award from the Miami University senior class, and in 1992 he received an Effective Educator award from the Richard T. Farmer School of Business Administration. Together with Richard T. O'Connell, Professor Bowerman has written 25 textbooks. These include *Forecasting, Time Series, and Regression: An Applied Approach* (also coauthored with Anne B. Koehler); *Linear Statistical Models: An Applied Approach*; *Regression Analysis: Unified Concepts, Practical Applications, and Computer Implementation* (also coauthored with Emily S. Murphree); and *Experimental Design: Unified Concepts, Practical Applications, and Computer Implementation* (also coauthored with Emily S. Murphree). The first edition of *Forecasting and Time Series* earned an Outstanding Academic Book award from *Choice* magazine. Professor Bowerman has also published a number of articles in applied stochastic process, time series forecasting, and statistical education. In his spare time, Professor Bowerman enjoys watching movies and sports, playing tennis, and designing houses.

Courtesy of Anne Drougas

Anne Drougas Anne M. Drougas is a Professor of Finance and Quantitative Methods at Dominican University in River Forest, Illinois. Over the course of her academic career, she has received three teaching awards and has developed and taught online and hybrid business statistics and finance courses. Her research is primarily in the areas of corporate finance, simulation, and business analytics with publications in a number of journals including the *Journal of Financial Education* and *Journal of Applied Business and Economics*. She spends her spare time with her family and serving on the board of directors for Hephzibah House, a social service agency for children in Oak Park, Illinois.

Courtesy of William Duckworth

William Duckworth William M. Duckworth specializes in statistics education and business applications of statistics. His professional affiliations have included the American Statistical Association (ASA), the International Association for Statistical Education (IASE), and the Decision Sciences Institute (DSI). Dr. Duckworth was also a member of the Undergraduate Statistics Education Initiative (USEI), which developed curriculum guidelines for undergraduate programs in statistical science that were officially adopted by the ASA. Dr. Duckworth has published research papers, been an invited speaker at professional meetings, and taught company training workshops, in addition to providing consulting and expert witness services to a variety of companies. During his tenure in the Department of Statistics at Iowa State University, his main responsibility was coordinating, teaching, and improving introductory business statistics courses. Dr. Duckworth currently teaches business analytics to both undergraduate and graduate students in the Heider College of Business at Creighton University.

Courtesy of Amy Froelich

Amy Froelich Amy G. Froelich received her Ph.D. in Statistics from the University of Illinois, Urbana-Champaign, and currently is Associate Professor and Director of Undergraduate Education in the Department of Statistics at Iowa State University. A specialist in undergraduate statistics education, she has taught over 2,700 students at Iowa State in the last 18 years, primarily in introductory statistics, probability and mathematical statistics, and categorical data analysis. Her research in statistics education and psychometrics and educational measurement has appeared in *The American Statistician*, the *Journal of Statistics Education, Teaching Statistics*, and the *Journal of Educational Measurement*, and she and her colleagues have received research funding from the National Science Foundation, the U.S. Department of Agriculture, and the U.S. Department of Education. Dr. Froelich has received several teaching and advising awards at Iowa State University and was the 2010 recipient of the Waller Education Award from the American Statistical Association. When not working, she enjoys reading, spending time with her family, and supporting her daughters' extracurricular activities.

Courtesy of Steve Muir/SAS

Ruth Hummel Ruth M. Hummel is an Academic Ambassador with JMP, a division of SAS specializing in desktop software for dynamic data visualization and analysis. As a technical advocate for the use of JMP® in academic settings, she supports professors and instructors who use JMP for teaching and research. She has been teaching and consulting since 2002, when she started her career as a high school math teacher. She has taught high school, undergraduate, and graduate courses in mathematics and statistics, and directed statistical research and analysis in a variety of fields. Ruth holds a Ph.D. in statistics from the Pennsylvania State University.

Courtesy of Kyle Moninger

Kyle Moninger Kyle B. Moninger instructs the Quantitative Business Curriculum at Bowling Green State University in Bowling Green, Ohio. He teaches and plans undergraduate courses in statistics and business calculus, serves on the Quantitative Business Curriculum committee, and supervises the college's math and statistics tutoring center. Kyle has been a visiting instructor three times at Tianjin Polytechnic University in Tianjin, China, and was previously a data scientist at Owens Corning in Toledo, Ohio, where he designed and implemented a corporate training program on business intelligence and analytics.

Courtesy of Pat Schur

Pat Schur Patrick J. Schur is a Senior Clinical Professor in the Department of Information Systems and Analytics in the Farmer School of Business at Miami University in Oxford, Ohio. He received his master's degree in statistics from Purdue University. He has been at Miami University for 11 years, teaching introductory statistics courses and advanced statistics courses including regression modeling, time series modeling, design of experiments, and statistical process control. Before joining Miami University, he worked at Procter & Gamble as a statistical consultant and also worked with multiple startup companies cutting across multiple industries.

AUTHORS' PREVIEW

Business Statistics and Analytics in Practice, Ninth Edition, provides a unique and flexible framework for teaching the introductory course in business statistics. This framework consists of

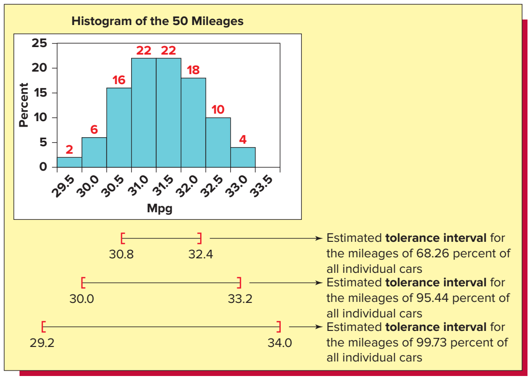
- A complete presentation of traditional business statistics, with improved discussions of introductory concepts, probability modeling, classical statistical inference (including a much clearer explanation of hypothesis testing), and regression and time series modeling.
- A complete presentation of business analytics, with topic coverage in six optional sections and two optional chapters: a section in Chapter 1 introducing analytics, five sections in Chapters 2 and 3 discussing descriptive analytics, and Chapters 5 and 16 discussing predictive analytics.
- Continuing case studies that facilitate student learning by presenting new concepts in the context of familiar situations.
- Business improvement conclusions—highlighted in yellow and designated by icons BI in the page margins—that explicitly show how statistical analysis leads to practical business decisions.
- Many new exercises.
- Use of Excel (including the Excel add-in MegaStat), Minitab, and JMP to carry out traditional statistical analysis. Use of JMP (and Excel and Minitab where possible) to carry out descriptive and predictive analytics.

We now discuss how these features are implemented in the book's 20 chapters.

Chapters 1, 2, and 3: Introductory concepts. Graphical and numerical descriptive methods. In an improved and simpler Chapter 1 we discuss data, variables, populations, and how to select random and other types of samples. Three case studies—**The Cell Phone Case, The Marketing Research Case,** and **The Car Mileage Case**—are used to illustrate sampling and how samples can be used to make statistical inferences.

In Chapters 2 and 3 we begin to formally discuss the statistical analysis used in making statistical inferences. For example, in Chapter 2 (graphical descriptive methods) we show how to construct a histogram of the car mileages that were sampled in **The Car Mileage Case** of Chapter 1. In Chapter 3 (numerical descriptive methods) we then use this histogram to help explain the Empirical Rule. As illustrated in Figure 3.15, this rule gives tolerance intervals providing estimates of the "lowest" and "highest" mileages that a new midsize car model should be expected to get in combined city and highway driving.

FIGURE 3.15 Estimated Tolerance Intervals in the Car Mileage Case



Estimated **tolerance interval** for the mileages of 68.26 percent of all individual cars

Estimated **tolerance interval** for the mileages of 95.44 percent of all individual cars

Estimated **tolerance interval** for the mileages of 99.73 percent of all individual cars

Chapters 1, 2, and 3: Six optional sections introducing business analytics and data mining and discussing descriptive analytics. In an optional section of Chapter 1 **The Disney Parks Case** introduces how business analytics and data mining are used to analyze big data. This case is then used in an optional section of Chapter 2 to help begin the book's discussion of descriptive analytics. Here, the optional section of Chapter 2 discusses what we call *graphical descriptive analytics,* and four optional sections in Chapter 3 (Part 2 of Chapter 3) discuss what we call *numerical descriptive analytics*. Included in the discussion of graphical descriptive analytics are gauges and dashboards (see Figure 2.35), bullet graphs and treemaps (see the Disney examples in Figures 2.36 and 2.37), and sparklines and data drill-down graphics. Included in the discussion of numerical descriptive analytics are association rules (see Figure 3.25), text mining (see Figure 3.27), hierarchical and k-means cluster analysis (see Figures 3.38 and 3.40), multidimensional scaling (which is part of the cluster analysis section), and factor analysis.

We believe that an early introduction to descriptive analytics will make statistics seem more useful and

FIGURE 2.35 A Dashboard of the Key Performance Indicators for an Airline

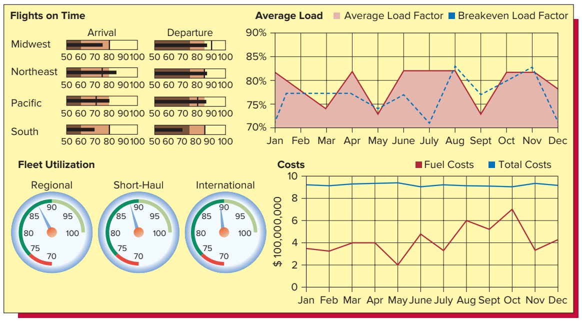


FIGURE 2.36 Excel Output of a Bullet Graph of Disney's Predicted Waiting Times (in minutes) for the Seven Epcot Rides Posted at 3 P.M. on February 21, 2015 DS DisneyTimes

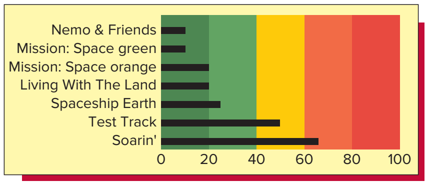


FIGURE 2.37 The Number of Ratings and the Mean Rating for Each of Seven Rides at Epcot (0 = Poor, 1 = Fair, 2 = Good, 3 = Very Good, 4 = Excellent, 5 = Superb) and a JMP Output of a Treemap of the Numbers of Ratings and the Mean Ratings

(b) JMP output of the treemap

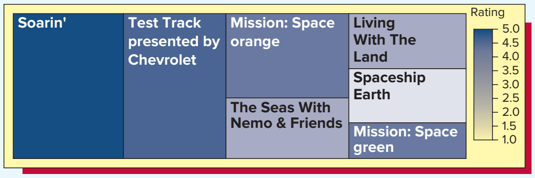
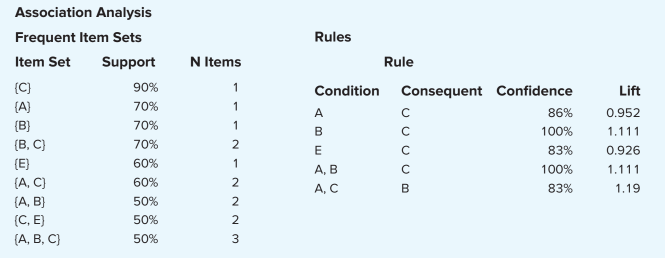


FIGURE 3.25 The JMP Output of an Association Rule Analysis of the DVD Renters Data DS DVDRent



Association Analysis

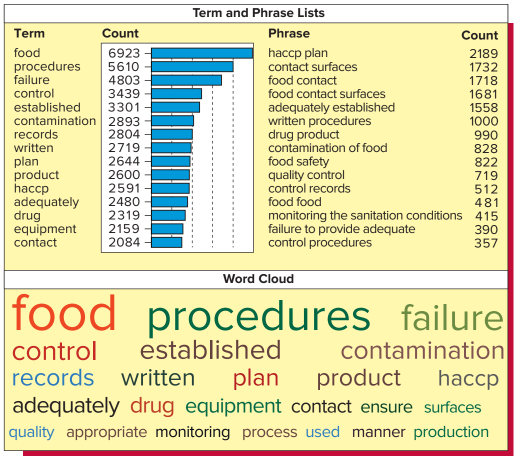
Frequent Item Sets

Item Set	Support	N Items
{C}	90%	1
{A}	70%	1
{B}	70%	1
{B, C}	70%	2
{E}	60%	1
{A, C}	60%	2
{A, B}	50%	2
{C, E}	50%	2
{A, B, C}	50%	3

Rules

Rule			
Condition	Consequent	Confidence	Lift
A	C	86%	0.952
B	C	100%	1.111
E	C	83%	0.926
A, B	C	100%	1.111
A, C	B	83%	1.19

FIGURE 3.27 The JMP Output of Part of a Term and Phrase List and Part of a Word Cloud in the FDA Citations Example DS FDACitat



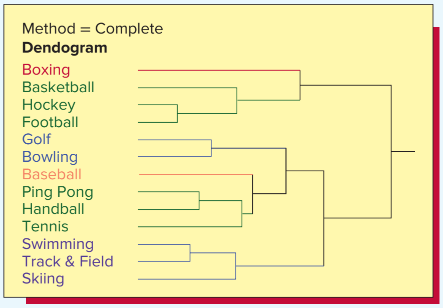
Term and Phrase Lists

Term	Count	Phrase	Count
food	6923	haccp plan	2189
procedures	5610	contact surfaces	1732
failure	4803	food contact	1718
control	3439	food contact surfaces	1681
established	3301	adequately established	1558
contamination	2893	written procedures	1000
records	2804	drug product	990
written	2719	contamination of food	828
plan	2644	food safety	822
product	2600	quality control	719
haccp	2591	control records	512
adequately	2480	food food	481
drug	2319	monitoring the sanitation conditions	415
equipment	2159	failure to provide adequate	390
contact	2084	control procedures	357

Word Cloud

food procedures failure control established contamination records written plan product haccp adequately drug equipment contact ensure surfaces quality appropriate monitoring process used manner production

FIGURE 3.38 A Second JMP Output of Hierarchical Clustering of the Sports Perception Data



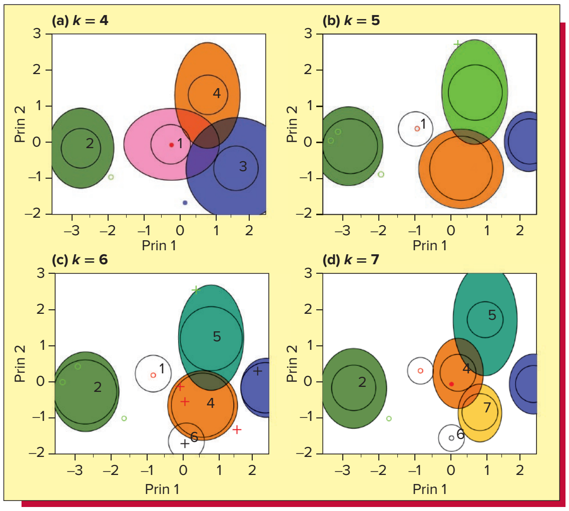
Method = Complete

Dendogram

Cluster Means

Cluster	Count	Fast/Slow	Comp/Simple	Team/Indv.	Easy/Hard	Nonc./Contact	Opponent/ Standard
1	1	3.07000	4.62000	6.62000	4.78000	6.02000	1.73000
2	3	1.99000	3.25333	1.60667	4.62000	5.77333	2.36333
3	2	5.60000	4.82500	5.99000	3.47500	1.71000	3.92000
4	1	4.78000	4.18000	2.16000	3.33000	3.60000	2.67000
5	3	2.86667	4.52667	5.21000	3.57000	2.32667	2.31000
6	3	2.60667	4.66667	5.24000	4.23333	2.54000	4.29667

FIGURE 3.40 The JMP Output of the Biplots in *k*-Means Clustering of the Sports Perception Data for *k* = 4, 5, 6, and 7


(a) *k* = 4

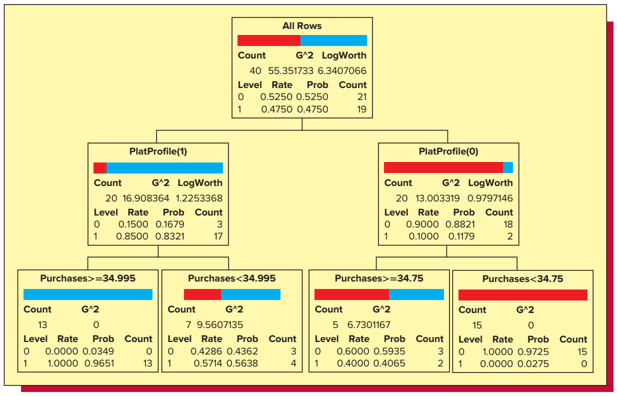
(b) *k* = 5

(c) *k* = 6

(d) *k* = 7

FIGURE 5.1 A JMP Classification Tree for the Card Upgrade Data DS CardUpgrade

(a) A JMP classification tree



All Rows
Count 40 G^2 55.351733 LogWorth 6.3407066

Level	Rate	Prob	Count
0	0.5250	0.5250	21
1	0.4750	0.4750	19

PlatProfile(1)
Count 20 G^2 16.908364 LogWorth 1.2253368

Level	Rate	Prob	Count
0	0.1500	0.1679	3
1	0.8500	0.8321	17

PlatProfile(0)
Count 20 G^2 13.003319 LogWorth 0.9797146

Level	Rate	Prob	Count
0	0.9000	0.8821	18
1	0.1000	0.1179	2

Purchases>=34.995
Count 13 G^2 0

Level	Rate	Prob	Count
0	0.0000	0.0349	0
1	1.0000	0.9651	13

Purchases<34.995
Count 7 G^2 9.5607135

Level	Rate	Prob	Count
0	0.4286	0.4362	3
1	0.5714	0.5638	4

Purchases>=34.75
Count 5 G^2 6.7301167

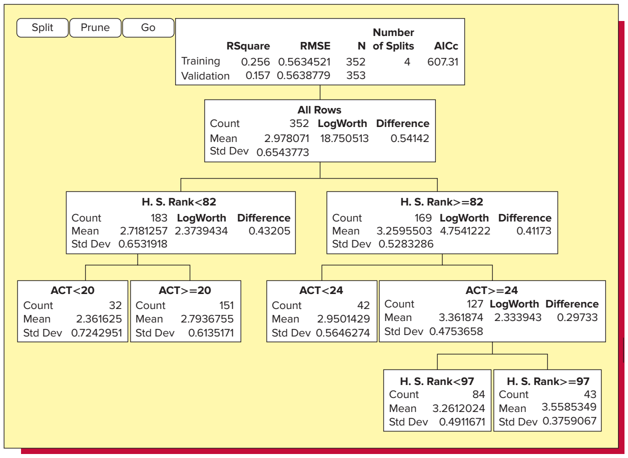
Level	Rate	Prob	Count
0	0.6000	0.5935	3
1	0.4000	0.4065	2

Purchases<34.75
Count 15 G^2 0

Level	Rate	Prob	Count
0	1.0000	0.9725	15
1	0.0000	0.0275	0

FIGURE 5.17 JMP Outputs of Building the Regression Tree in the GPA Example

(c) All four splits and the final regression tree



	RSquare	RMSE	N	Number of Splits	AICc
Training	0.256	0.5634521	352	4	607.31
Validation	0.157	0.5638779	353		

All Rows
Count 352 LogWorth 18.750513 Difference 0.54142
Mean 2.978071
Std Dev 0.6543773

H. S. Rank<82
Count 183 LogWorth 2.3739434 Difference 0.43205
Mean 2.7181257
Std Dev 0.6531918

H. S. Rank>=82
Count 169 LogWorth 4.7541222 Difference 0.41173
Mean 3.2595503
Std Dev 0.5283286

ACT<20
Count 32
Mean 2.361625
Std Dev 0.7242951

ACT>=20
Count 151
Mean 2.7936755
Std Dev 0.6135171

ACT<24
Count 42
Mean 2.9501429
Std Dev 0.5646274

ACT>=24
Count 127 LogWorth 2.333943 Difference 0.29733
Mean 3.361874
Std Dev 0.4753658

H. S. Rank<97
Count 84
Mean 3.2612024
Std Dev 0.4911671

H. S. Rank>=97
Count 43
Mean 3.5585349
Std Dev 0.3759067

FIGURE 5.28 Misclassification Rates in the k-Nearest Neighbors Card Upgrade Example

Training Set

K	Count	Misclassification Rate	Misclassifications
1	40	0.20000	8
2	40	0.12500	5*
3	40	0.20000	8
4	40	0.15000	6
5	40	0.17500	7
6	40	0.15000	6
7	40	0.15000	6
8	40	0.15000	6
9	40	0.15000	6
10	40	0.12500	5

Confusion Matrix for Best K=2

Training Set

Actual Upgrade	Predicted Count 0	1
0	18	3
1	2	17

FIGURE 5.36 Misclassifications in the Naive Bayes' Card Upgrade Example

Training Set

Count	Misclassification Rate	Misclassifications
40	0.12500	5

Confusion Matrix

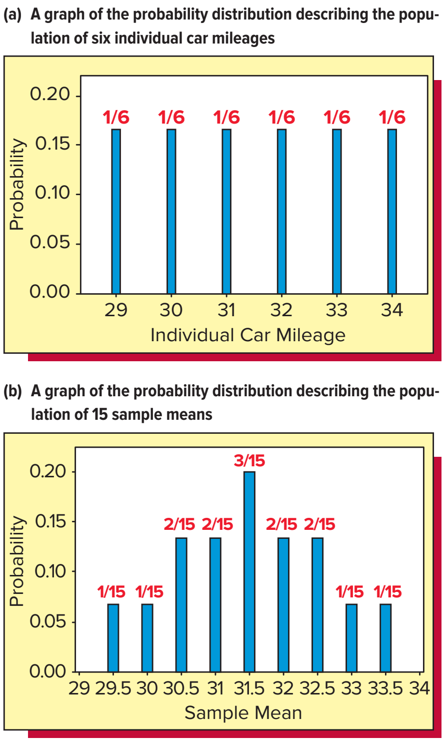
Training Set

Actual Upgrade	Predicted Count 0	1
0	17	4
1	1	18

relevant from the beginning and thus motivate students to be more interested in the entire course. However, our presentation gives instructors various choices. This is because, after covering the optional introduction to business analytics in Chapter 1, the five optional sections on descriptive analytics in Chapters 2 and 3 can be covered in any order without loss of continuity. Therefore, the instructor can choose which of the six optional business analytics sections to cover early, as part of the main flow of Chapters 1–3, and which to discuss later—perhaps with the predictive analytics discussed in Chapters 5 and 16. For courses with limited time to spend on descriptive analytics, we might recommend covering graphical descriptive analytics, association rules, and text mining. These topics are both very useful and easy to understand.

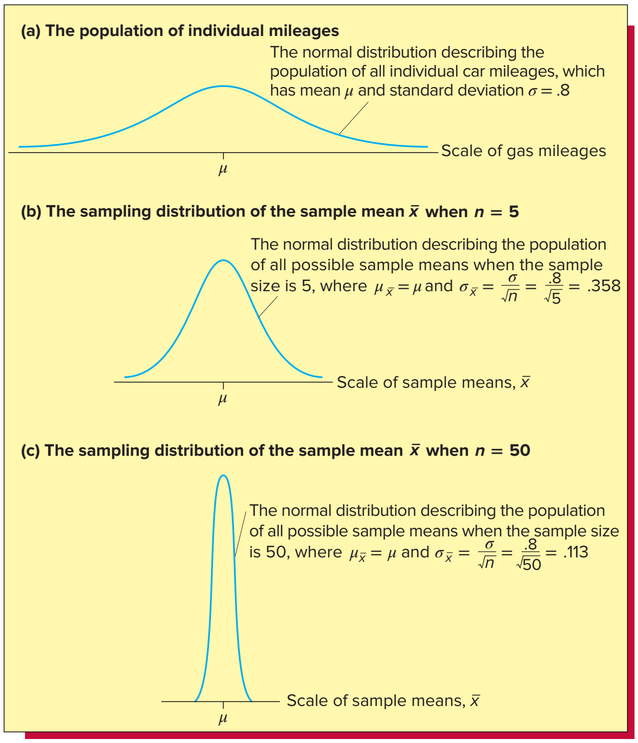
Chapters 4 and 5: Probability and probability modeling. Predictive analytics I (optional). Chapter 4 discusses probability and probability modeling by using motivating examples—**The Crystal Cable Case** and a real-world example of gender discrimination at a pharmaceutical company—to illustrate the probability rules. Optional Chapter 5 then uses the probability concepts of Chapter 4 and the descriptive statistics of Chapters 2 and 3 to discuss four predictive analytics: classification trees (see Figure 5.1), regression trees (see Figure 5.17). k-nearest neighbors (see Figure 5.28), and naive Bayes' classification (see Figure 5.36). These predictive analytics are called *nonparametric predictive analytics* and differ from the *parametric predictive analytics* discussed in Chapter 16. Parametric predictive analytics make predictions by using parametric equations that are evaluated by using the statistical inference techniques of Chapters 6 through 15. Nonparametric predictive analytics make predictions without using such equations and can be understood (from an applied standpoint) with a background of only descriptive statistics and probability. Chapters 5 and 16 are independent of each other and of the descriptive analytics sections in Chapters 2 and 3. Therefore, the instructor has the option to try to motivate student interest by covering Chapter 5 early, in the

FIGURE 8.1 A Comparison of Individual Car Mileages and Sample Means


(a) A graph of the probability distribution describing the population of six individual car mileages

(b) A graph of the probability distribution describing the population of 15 sample means

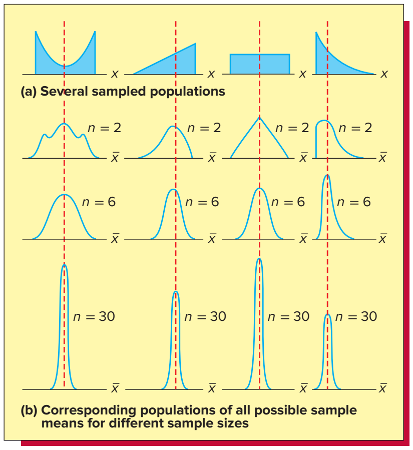
FIGURE 8.3 A Comparison of (1) the Population of All Individual Car Mileages, (2) the Sampling Distribution of the Sample Mean $\bar{x}$ When $n = 5$, and (3) the Sampling Distribution of the Sample Mean $\bar{x}$ When $n = 50$


(a) The population of individual mileages

(b) The sampling distribution of the sample mean $\bar{x}$ when $n = 5$

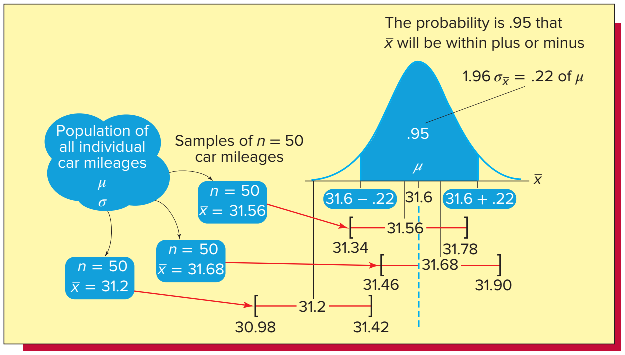
(c) The sampling distribution of the sample mean $\bar{x}$ when $n = 50$

FIGURE 8.5 The Central Limit Theorem Says That the Larger the Sample Size Is, the More Nearly Normally Distributed Is the Population of All Possible Sample Means


(a) Several sampled populations

(b) Corresponding populations of all possible sample means for different sample sizes

FIGURE 9.2 Three 95 Percent Confidence Intervals for μ


The probability is .95 that $\bar{x}$ will be within plus or minus $1.96\,\sigma_{\bar{x}} = .22$ of μ
.95
μ
$\bar{x}$
Population of all individual car mileages
μ
σ
Samples of $n = 50$ car mileages
$n = 50$
$\bar{x} = 31.56$
$n = 50$
$\bar{x} = 31.68$
$n = 50$
$\bar{x} = 31.2$
31.6 − .22
31.6
31.6 + .22
31.56
31.34
31.78
31.68
31.46
31.90
31.2
30.98
31.42

main flow of the course, or wait to cover Chapter 5 until later, perhaps with (before or after) the parametric predictive analytics in Chapter 16. For courses with limited time to spend on nonparametric predictive analytics, we might suggest covering just classification trees and regression trees.

Chapters 6–9: Discrete and continuous probability distributions. Sampling distributions and confidence intervals. Chapters 6 and 7 give discussions of discrete and continuous probability distributions (models) and feature practical examples illustrating the "rare event approach" to making a statistical inference. In Chapter 8, **The Car Mileage Case** is used to introduce sampling distributions and motivate the Central Limit Theorem (see Figures 8.1, 8.3, and 8.5). In Chapter 9, the automaker in **The Car Mileage Case** uses a confidence interval procedure specified by the Environmental Protection Agency (EPA) to find the EPA estimate of a new midsize model's true mean mileage and determine if the new midsize model deserves a federal tax credit (see Figure 9.2).

FIGURE 10.1 Testing H_0: $\mu = 50$ versus H_a: $\mu > 50$ by setting α Equal to .05

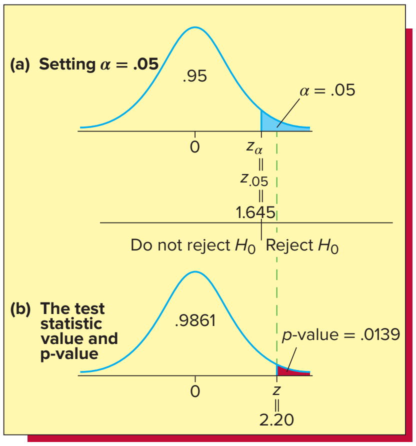


FIGURE 10.2 Testing H_0: $\mu = 50$ versus H_a: $\mu > 50$ by setting α Equal to .01

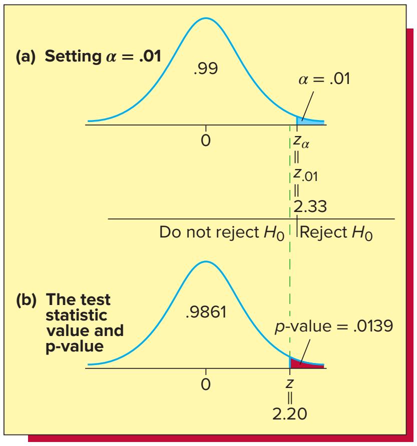


FIGURE 10.3 Testing H_0: $\mu = 19.5$ versus H_a: $\mu < 19.5$ by Using a Critical Value and a *p*-Value

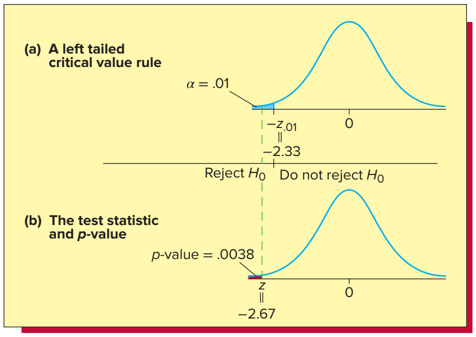
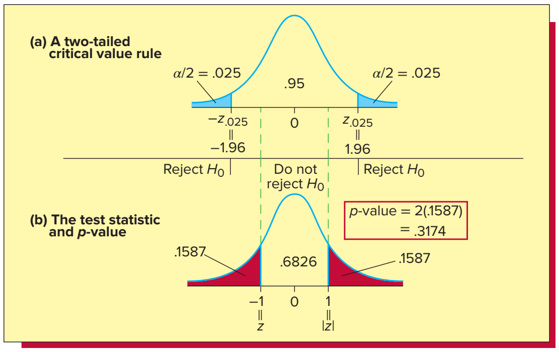


FIGURE 10.4 Testing H_0: $\mu = 330$ versus H_a: $\mu \neq 330$ by Using Critical Values and the *p*-Value



Chapters 10–13: Hypothesis testing. Two-sample procedures. Experimental design and analysis of variance. Chi-square tests. Chapter 10 discusses hypothesis testing and begins with a new section on formulating statistical hypotheses and the meanings of Type I and Type II errors. Three case studies—**The Trash Bag Case, The e-Billing Case,** and **The Valentine's Day Chocolate Case**—are then used in the next section to give a more unified and clearer discussion of the critical value rule and *p*-value approaches to performing a *z* test about the population mean. Specifically, for each type of alternative hypothesis, this discussion first illustrates the appropriate critical value rule in a graphical figure and then, in the same graphical figure, shows the appropriate *p*-value and explains why it is the more informative way to carry out the hypothesis test.

For example, the above Figures 10.1 and 10.2 are presented side-by-side in the text and illustrate testing a "greater than" alternative hypothesis in **The Trash Bag Case.** These figures show the different α's specified by two television networks evaluating a trash bag advertising claim, the different critical values that would have table looked up by an hypothesis tester using the critical value rule approach, and have the *p*-value immediately tells the hypothesis tester the results of the hypothesis test for any and all values of α. Similarly, Figures 10.3 and 10.4 illustrate the appropriate critical value rules and *p*-values for testing "less than" and "not equal to" alternative hypotheses.

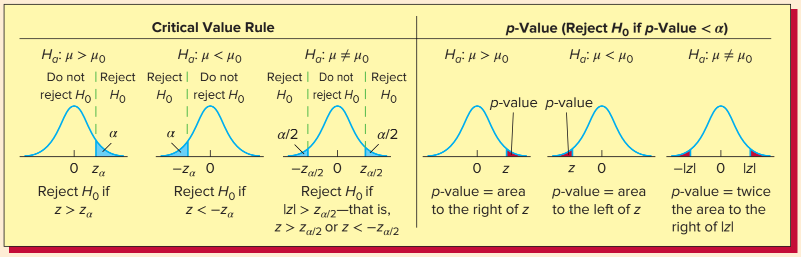
In addition, as the case studies are used to illustrate hypothesis testing, the *z* test about a population mean summary box and the five-step hypothesis testing procedure shown in the upper portion of the next page are developed. Here, although the true value of the population standard deviation is rarely known, the *z* test about a population mean summary box serves as an easily modifiable model for the book's other more practically useful hypothesis testing summary boxes—for example, for the *t* test about a population mean summary box and the *z* test about a population proportion summary box shown in the lower portion of the next page. Moreover, the five-step hypothesis testing procedure emphasizes that to successfully use a hypothesis testing summary box, we simply identify the alternative hypothesis being tested and then looking the summary box for the appropriate critical value rule and/or *p*-value.

Testing a Hypothesis about a Population Mean When σ Is Known

Null Hypothesis $H_0: \mu = \mu_0$

Test Statistic $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Assumptions Normal population or Large sample size



Critical Value Rule			p-Value (Reject H_0 if p-Value $< \alpha$)		
$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$	$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$	$H_a: \mu \neq \mu_0$
Reject H_0 if $z > z_\alpha$	Reject H_0 if $z < -z_\alpha$	Reject H_0 if $\lvert z\rvert > z_{\alpha/2}$—that is, $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	p-value = area to the right of z	p-value = area to the left of z	p-value = twice the area to the right of |z|

The Five Steps of Hypothesis Testing

1 State the null hypothesis H_0 and the alternative hypothesis H_a.

2 Specify the level of significance α.

3 Plan the sampling procedure and select the test statistic.

Using a critical value rule:

4 Use the summary box to find the critical value rule corresponding to the alternative hypothesis.

5 Collect the sample data, compute the value of the test statistic. and decide whether to reject H_0 by using the critical value rule. Interpret the statistical results.

Using a *p*-value rule:

4 Collect the sample data and compute the value of the test statistic.

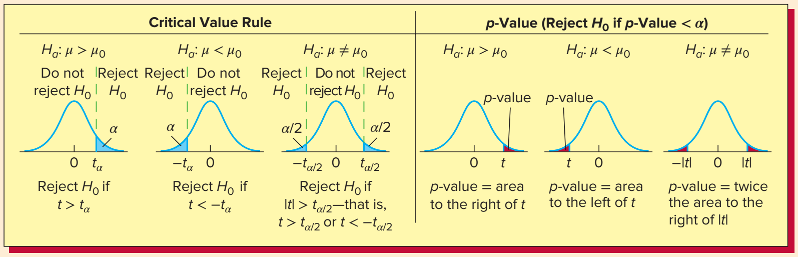
5 Use the summary box to find the *p*-value corresponding to the alternative hypothesis. Use the computed test statistic value to compute the *p*-value. Reject H_0 at level of significance α if the *p*-value is less than α. Interpret the statistical results.

A *t* Test about a Population Mean: σ Unknown

Null Hypothesis $H_0: \mu = \mu_0$

Test Statistic $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ $\quad df = n - 1$

Assumptions Normal population or Large sample size



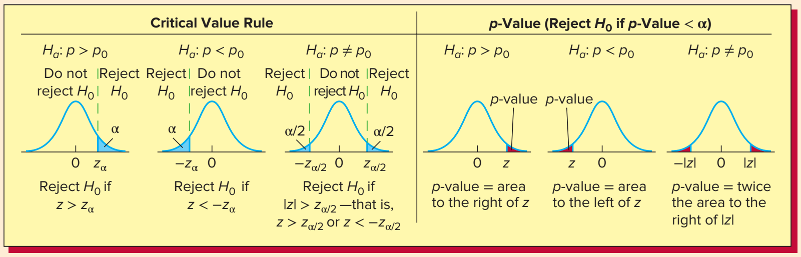
Critical Value Rule			p-Value (Reject H_0 if p-Value $< \alpha$)		
$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$	$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$	$H_a: \mu \neq \mu_0$
Reject H_0 if $t > t_\alpha$	Reject H_0 if $t < -t_\alpha$	Reject H_0 if $\lvert t\rvert > t_{\alpha/2}$—that is, $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$	p-value = area to the right of t	p-value = area to the left of t	p-value = twice the area to the right of |t|

A Large Sample Test about a Population Proportion

Null Hypothesis $H_0: p = p_0$

Test Statistic $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

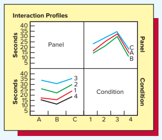
Assumptions[2] $np_0 \geq 5$ and $n(1 - p_0) \geq 5$



Critical Value Rule			p-Value (Reject H_0 if p-Value $< \alpha$)		
$H_a: p > p_0$	$H_a: p < p_0$	$H_a: p \neq p_0$	$H_a: p > p_0$	$H_a: p < p_0$	$H_a: p \neq p_0$
Reject H_0 if $z > z_\alpha$	Reject H_0 if $z < -z_\alpha$	Reject H_0 if $\lvert z\rvert > z_{\alpha/2}$ —that is, $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	p-value = area to the right of z	p-value = area to the left of z	p-value = twice the area to the right of |z|

FIGURE 12.12 The Display Panel Data and the JMP Output of a Two-Way ANOVA (for Exercise 12.21) DS Display2

	Emergency Condition			
Display Panel	1	2	3	4
A	17	25	31	14
	14	24	34	13
B	15	22	28	9
	12	19	31	10
C	21	29	32	15
	24	28	37	19



Least Squares Means Estimates

Panel	Estimate	Condition	Estimate
A	21.500000	1	17.166667
B	18.250000	2	24.500000
C	25.625000	3	32.166667
		4	13.333333

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	11	1482.4583	134.769	32.6713
Error	12	49.5000	4.125	Prob > F
C. Total	13	1531.9583		<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Panel	2	2	218.5833	26.4949	<.0001*
Condition	3	3	1247.4583	100.8047	<.0001*
Panel* Condition	6	6	16.4167	0.6633	0.6809

Tukey HSD All Pairwise Comparisons

Quantile = 2.66776, Adjusted DF = 12.0, Adjustment = Tukey

Panel	-Panel	Difference	Std Error	t Ratio	Prob > |t|	Lower 95%	Upper 95%
A	B	3.25000	1.015505	3.20	0.0194*	0.5409	5.95912
A	C	−4.12500	1.015505	−4.06	0.0042*	−6.8341	−1.41588
B	C	−7.37500	1.015505	−7.26	< .0001*	−10.0841	−4.66588

Tukey HSD All Pairwise Comparisons

Quantile = 2.9688, Adjusted DF = 12.0, Adjustment = Tukey

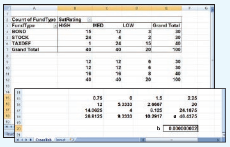
Condition	-Condition	Difference	Std Error	t Ratio	Prob > |t|	Lower 95%	Upper 95%
1	2	−7.3333	1.172604	−6.25	0.0002*	−10.8146	−3.8521
1	3	−15.0000	1.172604	−12.79	< .0001*	−18.4812	−11.5188
1	4	3.8333	1.172604	3.27	0.0297*	0.3521	7.3146
2	3	−7.6667	1.172604	−6.54	0.0001*	−11.1479	−4.1854
2	4	11.1667	1.172604	9.52	< .0001*	7.6854	14.6479
3	4	18.8333	1.172604	16.06	< .0001*	15.3521	22.3146

FIGURE 13.2 Partial Survey Results, and the Excel, Minitab, and JMP Outputs of a Contingency Table of Fund Type versus Level of Client Satisfaction DS Invest

(a) Survey results for the first three randomly selected clients

Client	Fund Type	Level of Satisfaction
1	BOND	HIGH
2	STOCK	HIGH
3	TAXDEF	MED

(b) The Excel output



```
Count of FundType  SatRating
FundType           HIGH     MED      LOW      Grand Total
BOND                 15      12        3       30
STOCK                24       4        2       30
TAXDEF                1      24       15       40
Grand Total          40      40       20      100

                     12      12        6       30
                     12      12        6       30
                     16      16        8       40
                     40      40       20      100

                   0.75       0      1.5       2.25
                     12  5.3333   2.6667       20
                14.0625       4    6.125       24.1875
                26.8125  9.3333  10.2917    a  46.4375

                                        b  0.000000002
```

a = Chi-square statistic.
b = *p*-value for chi-square.

Source: Microsoft Office Excel 2016

(c) The Minitab output

```
Rows: FundType     Columns: SatRating

            High     Med     Low      All
Bond          15      12       3       30
           50.00   40.00   10.00   100.00
           37.50   30.00   15.00    30.00
              12      12       6       30

Stock         24       4       2       30
           80.00   13.33    6.67   100.00
           60.00   10.00   10.00    30.00
              12      12       6       30

TaxDef         1      24      15       40
            2.50   60.00   37.50   100.00
            2.50   60.00   75.00    40.00
              16      16       8       40

All           40      40      20      100
           40.00   40.00   20.00   100.00
          100.00  100.00  100.00   100.00
              40      40      20      100

Pearson Chi-Square = 46.438, DF = 4
   P-Value = 0.000

Cell Contents:        Count
                      % of Row
                      % of Column
                      Expected count
```

(d) The JMP output

```
Contingency Table
                           SatRating
          Count      High      Med      Low     Total
          Total %
          Col %
          Row %
          Expected
          Bond         15       12        3        30
                    15.00    12.00     3.00     30.00
                    37.50    30.00    15.00
                    50.00    40.00    10.00
                       12       12        6
FundType  Stock        24        4        2        30
                    24.00     4.00     2.00     30.00
                    60.00    10.00    10.00
                    80.00    13.33     6.67
                       12       12        6
          TaxDef        1       24       15        40
                     1.00    24.00    15.00     40.00
                     2.50    60.00    75.00
                     2.50    60.00    37.50
                       16       16        8
          Total        40       40       20       100
                    40.00    40.00    20.00
          Tests
              N        DF    -LogLike    RSquare (U)
            100         4   27.699274         0.2626

          Test               ChiSquare      Prob>Chisq
          Likelihood Ratio      55.399         <.0001*
          Pearson               46.438         <.0001*
```

Hypothesis testing summary boxes are featured throughout Chapter 10, Chapter 11 (two-sample procedures), Chapter 12 (one-way, randomized block, and two-way analysis of variance), Chapter 13 (chi-square tests of goodness of fit and independence), and the remainder of the book. Furthermore, emphasis is placed throughout on assessing practical importance after testing for statistical significance. For example, as illustrated in Figure 12.12, if an *F* test finds a significant factor in an analysis of variance, we assess practical importance by finding point estimates of and confidence intervals for the differences in the effects of the different levels of the factor. As another example (see Figure 13.2), if a chi-square test rejects the hypothesis of independence between two variables, we assess practical importance by using the contingency table upon which the chi-square test is based to analyze the nature of the dependence between the variables.

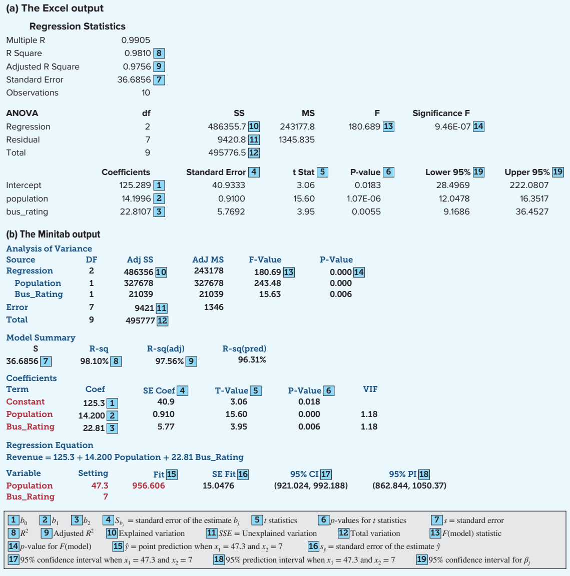
Chapters 14–17: Simple linear regression. Multiple regression and model building. Predictive analytics II (optional). Time series forecasting and index numbers. Chapter 14 discusses simple linear regression and illustrates the results of a simple linear regression analysis by using **The Tastee Sub Shop** (revenue prediction) **Case.** This same case is then used by the first seven sections of Chapter 15 (multiple regression and model building) to illustrate the results of a basic multiple regression analysis (see Figure 15.4). The last four sections of Chapter 15 continue the regression discussion by presenting four modeling topics that can be covered in any order without loss of continuity: dummy variables (including a discussion of interaction); quadratic variables and quantitative interaction variables; model building and the effects of multicollinearity (including model building for big data—see Figure 15.31); and residual analysis and diagnosing outlying and influential observations.

With the regression concepts of Chapters 14 and 15 as background, optional Chapter 16 extends these concepts and discusses three parametric predictive analytics: logistic regression (see Figure 16.5), linear discriminate analysis (see Figure 16.12), and neural networks (see Figures 16.17 and 16.19). Moreover, Chapter 17 extends the regression concepts in a different way and discusses time series forecasting methods, including an expanded presentation of exponential smoothing and a new and fuller (but understandable) presentation of the Box–Jenkins methodology.

Note that although we have used the term predictive analytics to refer only to the prediction methods of Chapters 5 and 16, the regression and time series methods of Chapters 14, 15, and 17 all predict (using a parametric equation) values of a response variable and thus are all (parametric) predictive analytics. We have used the term predictive analytics to refer only to the predictive methods of Chapters 5 and 16 because these methods are (for the most part) more modern methods that have been found to be particularly successful in analyzing big data. Together, the more classical parametric predictive analytics of Chapters 14, 15, and 17, along with the more modern nonparametric and parametric predictive analytics of Chapters 5 and 16 and the descriptive analytics of Chapters 2 and 3, make up a full second statistics course in business analytics.

Chapters 18–20: Concluding chapters. The book concludes with Chapters 18 (nonparametric statistics), Chapter 19 (decision theory), and website Chapter 20 (process improvement using control charts).

FIGURE 15.4 Excel and Minitab Outputs of a Regression Analysis of the Tasty Sub Shop Revenue Data in Table 15.1 Using the Model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



(a) The Excel output

Regression Statistics

Multiple R	0.9905
R Square	0.9810 $\boxed{8}$
Adjusted R Square	0.9756 $\boxed{9}$
Standard Error	36.6856 $\boxed{7}$
Observations	10

ANOVA	df	SS	MS	F	Significance F
Regression	2	486355.7 $\boxed{10}$	243177.8	180.689 $\boxed{13}$	9.46E-07 $\boxed{14}$
Residual	7	9420.8 $\boxed{11}$	1345.835		
Total	9	495776.5 $\boxed{12}$			

	Coefficients	Standard Error $\boxed{4}$	t Stat $\boxed{5}$	P-value $\boxed{6}$	Lower 95% $\boxed{19}$	Upper 95% $\boxed{19}$
Intercept	125.289 $\boxed{1}$	40.9333	3.06	0.0183	28.4969	222.0807
population	14.1996 $\boxed{2}$	0.9100	15.60	1.07E-06	12.0478	16.3517
bus_rating	22.8107 $\boxed{3}$	5.7692	3.95	0.0055	9.1686	36.4527

(b) The Minitab output

Analysis of Variance

Source	DF	Adj SS	AdJ MS	F-Value	P-Value
Regression	2	486356 $\boxed{10}$	243178	180.69 $\boxed{13}$	0.000 $\boxed{14}$
Population	1	327678	327678	243.48	0.000
Bus_Rating	1	21039	21039	15.63	0.006
Error	7	9421 $\boxed{11}$	1346		
Total	9	495777 $\boxed{12}$			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
36.6856 $\boxed{7}$	98.10% $\boxed{8}$	97.56% $\boxed{9}$	96.31%

Coefficients

Term	Coef	SE Coef $\boxed{4}$	T-Value $\boxed{5}$	P-Value $\boxed{6}$	VIF
Constant	125.3 $\boxed{1}$	40.9	3.06	0.018	
Population	14.200 $\boxed{2}$	0.910	15.60	0.000	1.18
Bus_Rating	22.81 $\boxed{3}$	5.77	3.95	0.006	1.18

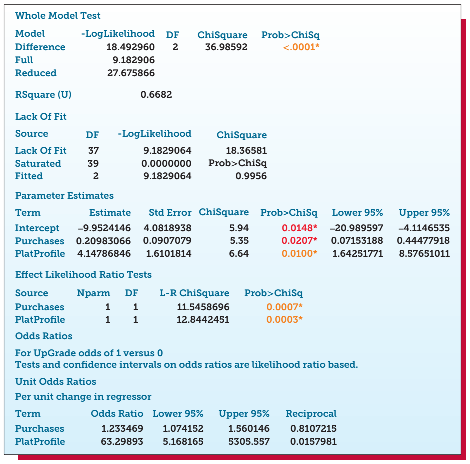
Regression Equation

Revenue = 125.3 + 14.200 Population + 22.81 Bus_Rating

Variable	Setting	Fit $\boxed{15}$	SE Fit $\boxed{16}$	95% CI $\boxed{17}$	95% PI $\boxed{18}$
Population	47.3	956.606	15.0476	(921.024, 992.188)	(862.844, 1050.37)
Bus_Rating	7				

$\boxed{1}$ b_0 $\boxed{2}$ b_1 $\boxed{3}$ b_2 $\boxed{4}$ S_{b_j} = standard error of the estimate b_j $\boxed{5}$ t statistics $\boxed{6}$ p-values for t statistics $\boxed{7}$ s = standard error
$\boxed{8}$ R^2 $\boxed{9}$ Adjusted R^2 $\boxed{10}$ Explained variation $\boxed{11}$ SSE = Unexplained variation $\boxed{12}$ Total variation $\boxed{13}$ F(model) statistic
$\boxed{14}$ p-value for F(model) $\boxed{15}$ $\hat{y}$ = point prediction when $x_1 = 47.3$ and $x_2 = 7$ $\boxed{16}$ $s_{\hat{y}}$ = standard error of the estimate $\hat{y}$
$\boxed{17}$ 95% confidence interval when $x_1 = 47.3$ and $x_2 = 7$ $\boxed{18}$ 95% prediction interval when $x_1 = 47.3$ and $x_2 = 7$ $\boxed{19}$ 95% confidence interval for β_j

FIGURE 16.5 JMP Output of a Logistic Regression of the Credit Card Upgrade Data



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	18.492960	2	36.98592	<.0001*
Full	9.182906			
Reduced	27.675866			

RSquare (U) 0.6682

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	37	9.1829064	18.36581
Saturated	39	0.0000000	Prob>ChiSq
Fitted	2	9.1829064	0.9956

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intercept	–9.9524146	4.0818938	5.94	0.0148*	–20.989597	–4.1146535
Purchases	0.20983066	0.0907079	5.35	0.0207*	0.07153188	0.44477918
PlatProfile	4.14786846	1.6101814	6.64	0.0100*	1.64251771	8.57651011

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Purchases	1	1	11.5458696	0.0007*
PlatProfile	1	1	12.8442451	0.0003*

Odds Ratios

For UpGrade odds of 1 versus 0
Tests and confidence intervals on odds ratios are likelihood ratio based.

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Purchases	1.233469	1.074152	1.560146	0.8107215
PlatProfile	63.29893	5.168165	5305.557	0.0157981

FIGURE 16.17 The Single Layer Perceptron

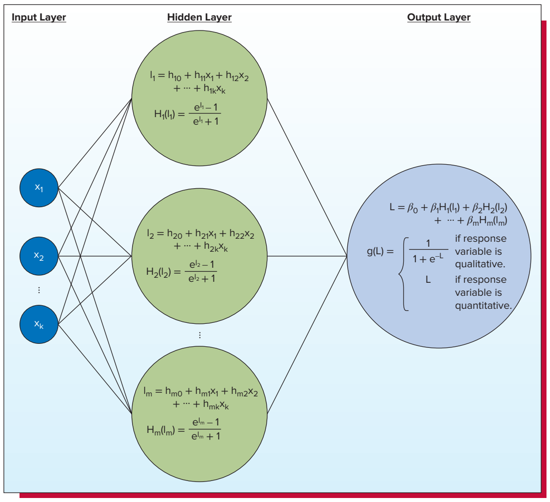
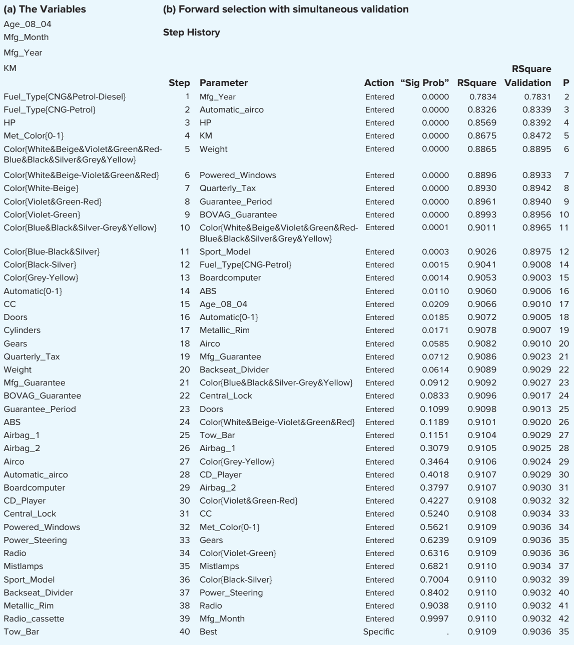

Input Layer
Hidden Layer
Output Layer
x_1
x_2
x_k
$l_1 = h_{10} + h_{11}x_1 + h_{12}x_2 + \cdots + h_{1k}x_k$
$H_1(l_1) = \frac{e^{l_1}-1}{e^{l_1}+1}$
$l_2 = h_{20} + h_{21}x_1 + h_{22}x_2 + \cdots + h_{2k}x_k$
$H_2(l_2) = \frac{e^{l_2}-1}{e^{l_2}+1}$
$l_m = h_{m0} + h_{m1}x_1 + h_{m2}x_2 + \cdots + h_{mk}x_k$
$H_m(l_m) = \frac{e^{l_m}-1}{e^{l_m}+1}$
$L = \beta_0 + \beta_1H_1(l_1) + \beta_2H_2(l_2) + \cdots + \beta_mH_m(l_m)$
$g(L) = \begin{cases} \frac{1}{1+e^{-L}} & \text{if response variable is qualitative.} \\ L & \text{if response variable is quantitative.} \end{cases}$

FIGURE 15.31 The JMP Output of the Potential Quantitative Independent Variables, Including the Dummy Variables, and Forward Selection with Simultaneous Validation in the Used Toyota Corolla Sales Price Example



(a) The Variables

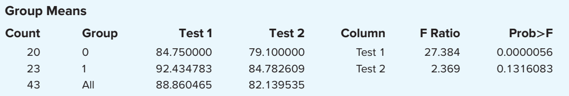
Age_08_04
Mfg_Month
Mfg_Year
KM
Fuel_Type{CNG&Petrol-Diesel}
Fuel_Type{CNG-Petrol}
HP
Met_Color{0-1}
Color{White&Beige&Violet&Green&Red-Blue&Black&Silver&Grey&Yellow}
Color{White&Beige-Violet&Green&Red}
Color{White-Beige}
Color{Violet&Green-Red}
Color{Violet-Green}
Color{Blue&Black&Silver-Grey&Yellow}
Color{Blue-Black&Silver}
Color{Black-Silver}
Color{Grey-Yellow}
Automatic{0-1}
CC
Doors
Cylinders
Gears
Quarterly_Tax
Weight
Mfg_Guarantee
BOVAG_Guarantee
Guarantee_Period
ABS
Airbag_1
Airbag_2
Airco
Automatic_airco
Boardcomputer
CD_Player
Central_Lock
Powered_Windows
Power_Steering
Radio
Mistlamps
Sport_Model
Backseat_Divider
Metallic_Rim
Radio_cassette
Tow_Bar

(b) Forward selection with simultaneous validation

Step History

Step	Parameter	Action	"Sig Prob"	RSquare	RSquare Validation	P
1	Mfg_Year	Entered	0.0000	0.7834	0.7831	2
2	Automatic_airco	Entered	0.0000	0.8326	0.8339	3
3	HP	Entered	0.0000	0.8569	0.8392	4
4	KM	Entered	0.0000	0.8675	0.8472	5
5	Weight	Entered	0.0000	0.8865	0.8895	6
6	Powered_Windows	Entered	0.0000	0.8896	0.8933	7
7	Quarterly_Tax	Entered	0.0000	0.8930	0.8942	8
8	Guarantee_Period	Entered	0.0000	0.8961	0.8940	9
9	BOVAG_Guarantee	Entered	0.0000	0.8993	0.8956	10
10	Color{White&Beige&Violet&Green&Red-Blue&Black&Silver&Grey&Yellow}	Entered	0.0001	0.9011	0.8965	11
11	Sport_Model	Entered	0.0003	0.9026	0.8975	12
12	Fuel_Type{CNG-Petrol}	Entered	0.0015	0.9041	0.9008	14
13	Boardcomputer	Entered	0.0014	0.9053	0.9003	15
14	ABS	Entered	0.0110	0.9060	0.9006	16
15	Age_08_04	Entered	0.0209	0.9066	0.9010	17
16	Automatic{0-1}	Entered	0.0185	0.9072	0.9005	18
17	Metallic_Rim	Entered	0.0171	0.9078	0.9007	19
18	Airco	Entered	0.0585	0.9082	0.9010	20
19	Mfg_Guarantee	Entered	0.0712	0.9086	0.9023	21
20	Backseat_Divider	Entered	0.0614	0.9089	0.9029	22
21	Color{Blue&Black&Silver-Grey&Yellow}	Entered	0.0912	0.9092	0.9027	23
22	Central_Lock	Entered	0.0833	0.9096	0.9017	24
23	Doors	Entered	0.1099	0.9098	0.9013	25
24	Color{White&Beige-Violet&Green&Red}	Entered	0.1189	0.9101	0.9020	26
25	Tow_Bar	Entered	0.1151	0.9104	0.9029	27
26	Airbag_1	Entered	0.3079	0.9105	0.9025	28
27	Color{Grey-Yellow}	Entered	0.3464	0.9106	0.9024	29
28	CD_Player	Entered	0.4018	0.9107	0.9029	30
29	Airbag_2	Entered	0.3797	0.9107	0.9030	31
30	Color{Violet&Green-Red}	Entered	0.4227	0.9108	0.9032	32
31	CC	Entered	0.5240	0.9108	0.9034	33
32	Met_Color{0-1}	Entered	0.5621	0.9109	0.9036	34
33	Gears	Entered	0.6239	0.9109	0.9036	35
34	Color{Violet-Green}	Entered	0.6316	0.9109	0.9036	36
35	Mistlamps	Entered	0.6821	0.9110	0.9034	37
36	Color{Black-Silver}	Entered	0.7004	0.9110	0.9032	39
37	Power_Steering	Entered	0.8402	0.9110	0.9032	40
38	Radio	Entered	0.9038	0.9110	0.9032	41
39	Mfg_Month	Entered	0.9997	0.9110	0.9032	42
40	Best	Specific	.	0.9109	0.9036	35

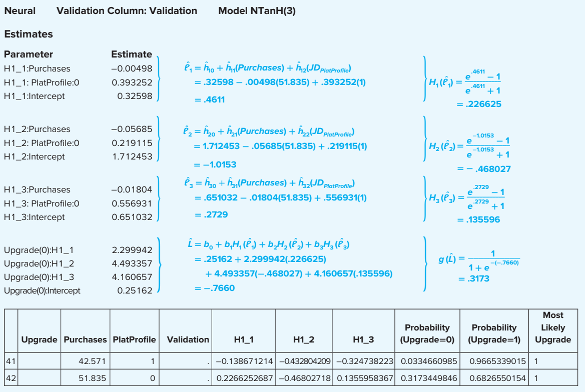
FIGURE 16.12 The Group Means and p-Values for Test 1 and Test 2



Group Means

Count	Group	Test 1	Test 2	Column	F Ratio	Prob>F
20	0	84.750000	79.100000	Test 1	27.384	0.0000056
23	1	92.434783	84.782609	Test 2	2.369	0.1316083
43	All	88.860465	82.139535			

FIGURE 16.19 JMP Output of Neural Network Estimation for the Credit Card Upgrade Data DS CardUpgrade



Neural Validation Column: Validation Model NTanH(3)

Estimates

Parameter	Estimate
H1_1:Purchases	–0.00498
H1_1: PlatProfile:0	0.393252
H1_1:Intercept	0.32598
H1_2:Purchases	–0.05685
H1_2: PlatProfile:0	0.219115
H1_2:Intercept	1.712453
H1_3:Purchases	–0.01804
H1_3: PlatProfile:0	0.556931
H1_3:Intercept	0.651032
Upgrade(0):H1_1	2.299942
Upgrade(0):H1_2	4.493357
Upgrade(0):H1_3	4.160657
Upgrade(0):Intercept	0.25162

$$\hat{\ell}_1 = \hat{h}_{10} + \hat{h}_{11}(Purchases) + \hat{h}_{12}(JD_{PlatProfile}) = .32598 - .00498(51.835) + .393252(1) = .4611$$

$$H_1(\hat{\ell}_1) = \frac{e^{.4611}-1}{e^{.4611}+1} = .226625$$

$$\hat{\ell}_2 = \hat{h}_{20} + \hat{h}_{21}(Purchases) + \hat{h}_{22}(JD_{PlatProfile}) = 1.712453 - .05685(51.835) + .219115(1) = -1.0153$$

$$H_2(\hat{\ell}_2) = \frac{e^{-1.0153}-1}{e^{-1.0153}+1} = -.468027$$

$$\hat{\ell}_3 = \hat{h}_{30} + \hat{h}_{31}(Purchases) + \hat{h}_{32}(JD_{PlatProfile}) = .651032 - .01804(51.835) + .556931(1) = .2729$$

$$H_3(\hat{\ell}_3) = \frac{e^{.2729}-1}{e^{.2729}+1} = .135596$$

$$\hat{L} = b_0 + b_1H_1(\hat{\ell}_1) + b_2H_2(\hat{\ell}_2) + b_3H_3(\hat{\ell}_3) = .25162 + 2.299942(.226625) + 4.493357(-.468027) + 4.160657(.135596) = -.7660$$

$$g(\hat{L}) = \frac{1}{1+e^{-(-.7660)}} = .3173$$

	Upgrade	Purchases	PlatProfile	Validation	H1_1	H1_2	H1_3	Probability (Upgrade=0)	Probability (Upgrade=1)	Most Likely Upgrade
41		42.571	1	.	–0.138671214	–0.432804209	–0.324738223	0.0334660985	0.9665339015	1
42		51.835	0	.	0.2266252687	–0.46802718	0.1355958367	0.3173449846	0.6826550154	1

Mc Graw Hill Education connect®

Students—study more efficiently, retain more and achieve better outcomes. Instructors—focus on what you love—teaching.

SUCCESSFUL SEMESTERS INCLUDE CONNECT

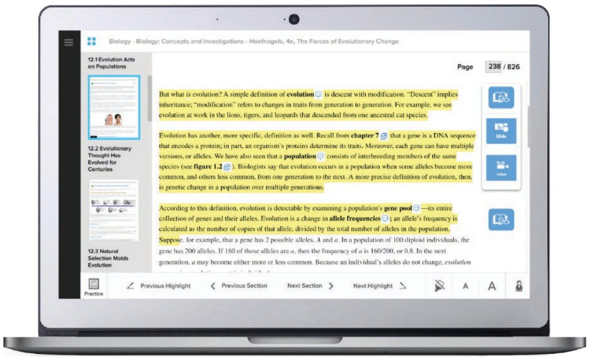
For Instructors

You're in the driver's seat.

Want to build your own course? No problem. Prefer to use our turnkey, prebuilt course? Easy. Want to make changes throughout the semester? Sure. And you'll save time with Connect's auto-grading too.

65%

Less Time Grading



Biology - Biology: Concepts and Investigations - Hoefnagels, 4e, The Forces of Evolutionary Change

12.1 Evolution Acts on Populations

12.2 Evolutionary Thought Has Evolved for Centuries

12.3 Natural Selection Molds Evolution

Page 238 / 826

But what *is* evolution? A simple definition of **evolution** is descent with modification. "Descent" implies inheritance; "modification" refers to changes in traits from generation to generation. For example, we see evolution at work in the lions, tigers, and leopards that descended from one ancestral cat species.

Evolution has another, more specific, definition as well. Recall from **chapter 7** that a gene is a DNA sequence that encodes a protein; in part, an organism's proteins determine its traits. Moreover, each gene can have multiple versions, or alleles. We have also seen that a **population** consists of interbreeding members of the same species (see **figure 1.2**). Biologists say that evolution occurs in a population when some alleles become more common, and others less common, from one generation to the next. A more precise definition of evolution, then, is genetic change in a population over multiple generations.

According to this definition, evolution is detectable by examining a population's **gene pool**—its entire collection of genes and their alleles. Evolution is a change in **allele frequencies**; an allele's frequency is calculated as the number of copies of that allele, divided by the total number of alleles in the population. Suppose, for example, that a gene has 2 possible alleles, *A* and *a*. In a population of 100 diploid individuals, the gene has 200 alleles. If 160 of those alleles are *a*, then the frequency of *a* is 160/200, or 0.8. In the next generation, *a* may become either more or less common. Because an individual's alleles do not change, *evolution*

They'll thank you for it.

Adaptive study resources like SmartBook® help your students be better prepared in less time. You can transform your class time from dull definitions to dynamic debates. Hear from your peers about the benefits of Connect at **www.mheducation.com/highered/connect**

Make it simple, make it affordable.

Connect makes it easy with seamless integration using any of the major Learning Management Systems—Blackboard®, Canvas, and D2L, among others—to let you organize your course in one convenient location. Give your students access to digital materials at a discount with our inclusive access program. Ask your McGraw-Hill representative for more information.

©Hill Street Studios/Tobin Rogers/Blend Images LLC

Solutions for your challenges.

A product isn't a solution. Real solutions are affordable, reliable, and come with training and ongoing support when you need it and how you want it. Our Customer Experience Group can also help you troubleshoot tech problems—although Connect's 99% uptime means you might not need to call them. See for yourself at **status.mheducation.com**

For Students

©Shutterstock/wavebreakmedia

Effective, efficient studying.

Connect helps you be more productive with your study time and get better grades using tools like SmartBook, which highlights key concepts and creates a personalized study plan. Connect sets you up for success, so you walk into class with confidence and walk out with better grades.

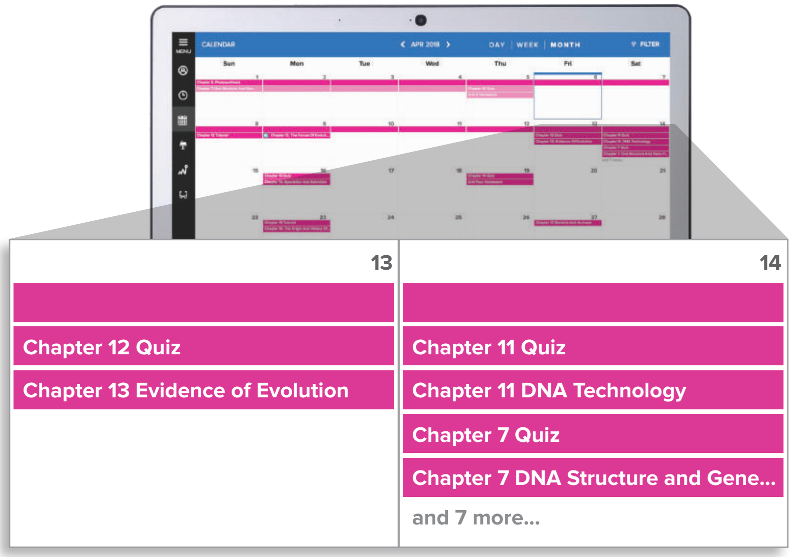
> “I really liked this app—it made it easy to study when you don't have your textbook in front of you.”
>
> - Jordan Cunningham,
> Eastern Washington University

Study anytime, anywhere.

Download the free ReadAnywhere app and access your online eBook when it’s convenient, even if you’re offline. And since the app automatically syncs with your eBook in Connect, all of your notes are available every time you open it. Find out more at **www.mheducation.com/readanywhere**

No surprises.

The Connect Calendar and Reports tools keep you on track with the work you need to get done and your assignment scores. Life gets busy; Connect tools help you keep learning through it all.



13

Chapter 12 Quiz

Chapter 13 Evidence of Evolution

14

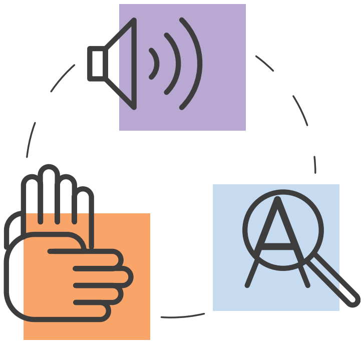
Chapter 11 Quiz

Chapter 11 DNA Technology

Chapter 7 Quiz

Chapter 7 DNA Structure and Gene...

and 7 more...



Learning for everyone.

McGraw-Hill works directly with Accessibility Services Departments and faculty to meet the learning needs of all students. Please contact your Accessibility Services office and ask them to email accessibility@mheducation.com, or visit **www.mheducation.com/about/accessibility.html** for more information.

ADDITIONAL RESOURCES

MEGASTAT® FOR MICROSOFT EXCEL® (AND EXCEL: MAC)

MegaStat is a full-featured Excel add-in by J. B. Orris of Butler University that is available with this text. It performs statistical analyses within an Excel workbook. It does basic functions such as descriptive statistics, frequency distributions, and probability calculations, as well as hypothesis testing, ANOVA, and regression.

MegaStat output is carefully formatted. Ease-of-use features include AutoExpand for quick data selection and Auto Label detect. Since MegaStat is easy to use, students can focus on learning statistics without being distracted by the software. MegaStat is always available from Excel's main menu. Selecting a menu item pops up a dialog box. MegaStat works with all recent versions of Excel. For more information, go to **mhhe.com/megastat**.

MINITAB®

Minitab® Student Version 18 is available to help students solve the business statistics exercises in the text. This software is available in the student version and can be packaged with any McGraw-Hill business statistics text.

JMP®

JMP Student Edition is an easy-to-use streamlined version of JMP software for both Windows and Mac that provides all the statistical analysis and graphical tools covered in introductory and many intermediate statistics courses.

CHAPTER-BY-CHAPTER REVISIONS FOR 9TH EDITION:

Chapter 1

- Improved and simpler introduction to business statistics
- Completely rewritten and much improved introduction to business analytics
- Appendix on using JMP added

Chapter 2

- JMP examples and exercises added
- Appendix on using JMP added

Chapter 3

- Improved discussion of association rules
- New section on text mining and latent semantic analysis added
- Much improved discussions of hierarchical clustering, *k*-means clustering, and multidimensional scaling; Improved discussion of factor analysis
- JMP examples and exercises added
- Appendix on using JMP added

Chapter 4

- Improved discussion of probability modeling

Chapter 5

- A new chapter on "nonparametric" prediction analytics—classification trees, regression trees, k-nearest neighbors, and naive Bayes' classification
- JMP examples and exercises added
- Appendix on using JMP added

Chapter 6 (formerly 5)

- JMP examples and exercises added
- Appendix on using JMP added

Chapter 7 (formerly 6)

- JMP examples and exercises added
- Appendix on using JMP added

Chapter 8 (formerly 7)

- No significant changes

Chapter 9 (formerly 8)

- JMP examples and exercises added
- Appendix on using JMP added

Chapter 10 (formerly 9)

- Improved section on formulating statistical hypotheses and the meanings of Type I and Type II errors
- Much improved (more unified, simpler, clearer, and shorter) explanation of using critical value rules and *p*-values to test hypotheses
- JMP examples and exercises added
- Appendix on using JMP added

Chapter 11 (formerly 10)

- JMP examples and exercises added
- Appendix on using JMP added

Chapter 12 (formerly 11)

- JMP examples and exercises added
- Appendix on using JMP added

Chapter 13 (formerly 12)

- JMP examples and exercises added
- Appendix on using JMP added

Chapter 14 (formerly 13)

- JMP examples and exercises added
- Appendix on using JMP added

Chapter 15 (formerly 14)

- New section on model building for big data added
- Improved discussion of diagnosing outlying and influential observations
- JMP examples and exercises added
- Appendix on using JMP added

Chapter 16

- A new chapter on "parametric" predictive analytics–logistic regression, linear discriminate analysis, and neural networks
- JMP examples and exercises added
- Appendix on using JMP added

Chapter 17 (formerly 15)

- Expanded discussion of exponential smoothing
- New and fuller (but understandable) discussion of the Box—Jenkins methodology
- JMP examples and exercises added
- Appendix on using JMP added

Chapter 18 (formerly 17)

- JMP examples and exercises added
- Appendix on using JMP added

Chapter 19 (formerly 18)

- No significant changes

Chapter 20 (on website–formerly 16)

- Appendix on using JMP added

ACKNOWLEDGMENTS

We wish to thank many people who have helped to make this book a reality. As indicated on the title page, we thank Professor Steven C. Huchendorf, University of Minnesota, and Dawn C. Porter, University of Southern California, for major contributions to this book. We also thank former co-author Emily Murphree for all of her excellent work on previous editions.

We wish to thank the people at McGraw-Hill for their dedication to this book. These people include senior brand manager Noelle Bathurst, who is extremely helpful to the authors; senior development editor Tobi Philips who has shown great dedication to the improvement of this book; content project manager Fran Simon, who has very capably and diligently guided this book through its production and who has been a tremendous help to the authors; and our former executive editor Steve Scheutz, who always greatly supported our books. We also thank lead product developer Michelle Janicek for her tremendous help in developing this new edition; our former executive editor Scott Isenberg for the tremendous help he has given us in developing all of our McGraw-Hill business statistics books; and our former executive editor Dick Hercher, who persuaded us to publish with McGraw-Hill.

We wish to thank Larry White, Eastern Illinois University, for revising the Test Bank and Vickie Fry, Indiana University Bloomington, for revising the PowerPoints and reviewing the Test Bank. Additional thanks to our co-authors for their work above and beyond the revision of the text: Patrick Schur, Miami University, for his work developing learning resources; Kyle Moninger, Bowling Green State University, for updating and revising the LearnSmart content; and Amy Froelich, Iowa State University, for developing the Solutions Manual. Most importantly, we wish to thank our families for their acceptance, unconditional love, and support.

DEDICATION

Bruce L. Bowerman

— Richard T. O'Connell, my best friend and wonderful co-author and person, whom I miss so much.

— Herbert T. David, my brilliant and kind Ph.D. advisor, who generously helped so many Ph.D. students.

— All of my loved ones.

Anne Drougas

I would like to dedicate this book to my parents, Arthur and Mary, and my sister, Cathy, who have been the greatest blessings in my life, my Ph.D. dissertation advisor, John, and Steve.

William Duckworth

To my wife, Shelia, and children—Billy, Kim, and Andrew: You have been supportive, helpful, and patient during this project and I thank you.

Amy Froelich

To my husband, Jim, and daughters, Sarah and Jamie.

To my Mom and Dad and brother, Scott.

Ruth Hummel:

To Bruce, for his friendship and the many interesting statistical and non-statistical conversations we've had.

Kyle Moninger

Dedicated to my family. They are everything to me.

Pat Schur

To my wife, children, and grandchildren

— Lorie

— Andy, Manda, and Angie

— Cooper, Chloe, Emma, Gracie, and Nolan

BRIEF CONTENTS

CONTENTS

Chapter 5

Predictive Analytics I: Trees, k-Nearest Neighbors, Naive Bayes', and Ensemble Estimates

Chapter 6

Discrete Random Variables

Chapter 7

Continuous Random Variables

Chapter 8

Sampling Distributions

Chapter 9

Confidence Intervals

Chapter 10

Hypothesis Testing

Chapter 11
Statistical Inferences Based on Two Samples

Chapter 12
Experimental Design and Analysis of Variance

Chapter 13
Chi-Square Tests

Chapter 14
Simple Linear Regression Analysis

Chapter 15
Multiple Regression and Model Building

Chapter 16
Predictive Analytics II: Logistic Regression, Discriminate Analysis, and Neural Networks

Chapter 17

Time Series Forecasting and Index Numbers

Chapter 18

Nonparametric Methods

Chapter 19

Decision Theory

Chapter 20 (Online)

Process Improvement Using Control Charts for Website

Appendix A

Appendix B (Online)

This page intentionally left blank

CHAPTER 1

An Introduction to Business Statistics and Analytics

©Tetra Images/Alamy

Learning Objectives

After mastering the material in this chapter, you will be able to:

LO1-1 Define a variable.

LO1-2 Describe the difference between a quantitative variable and a qualitative variable.

LO1-3 Describe the difference between cross-sectional data and time series data.

LO1-4 Construct and interpret a time series (runs) plot.

LO1-5 Identify the different types of data sources: existing data sources, experimental studies, and observational studies.

LO1-6 Explain the basic ideas of data warehousing and big data.

LO1-7 Describe the difference between a population and a sample.

LO1-8 Distinguish between descriptive statistics and statistical inference.

LO1-9 Explain the concept of random sampling and select a random sample.

LO1-10 Explain some of the uses of business analytics and data mining (Optional).

LO1-11 Identify the ratio, interval, ordinal, and nominative scales of measurement (Optional).

LO1-12 Describe the basic ideas of stratified random, cluster, and systematic sampling (Optional).

LO1-13 Describe basic types of survey questions, survey procedures, and sources of error (Optional).

Chapter Outline

The subject of statistics involves the study of how to collect, analyze, and interpret data. **Data are facts and figures from which conclusions can be drawn**. Such conclusions are important to the decision making of many professions and organizations. For example, **economists** use conclusions drawn from the latest data on unemployment and inflation to help the government make policy decisions. **Financial planners** use recent trends in stock market prices and economic conditions to make investment decisions. **Accountants** use **sample data** concerning a company's *actual sales revenues* to assess whether the company's *claimed sales revenues* are valid. **Marketing professionals** and **data miners** help businesses decide which products to develop and market and which consumers to target in marketing campaigns by using data that reveal consumer preferences. **Production supervisors** use manufacturing data to evaluate, control, and improve product quality. **Politicians** rely on data from public opinion polls to formulate legislation and to devise campaign strategies. **Physicians and hospitals** use data on the effectiveness of drugs and surgical procedures to provide patients with the best possible treatment.

In this chapter we begin to see how we collect and analyze data. As we proceed through the chapter, we introduce several case studies. These case studies (and others to be introduced later) are revisited throughout later chapters as we learn the statistical methods needed to analyze them. Briefly, we will begin to study four cases:

The Cell Phone Case: A bank estimates its cellular phone costs and decides whether to outsource management of its wireless resources by studying the calling patterns of its employees.

The Marketing Research Case: A beverage company investigates consumer reaction to a new bottle design for one of its popular soft drinks.

The Car Mileage Case: To determine if it qualifies for a federal tax credit based on fuel economy, an automaker studies the gas mileage of its new midsize model.

The Disney Parks Case: Walt Disney World Parks and Resorts in Orlando, Florida, manages Disney parks worldwide and uses data gathered from its guests to give these guests a more "magical" experience and increase Disney revenues and profits.

1.1 Data

LO1-1

Define a variable.

Data sets, elements, and variables

We have said that data are facts and figures from which conclusions can be drawn. Together, the data that are collected for a particular study are referred to as a **data set.** For example, Table 1.1 is a data set that gives information about the new homes sold in a Florida luxury home development over a recent three-month period. Potential home buyers could choose either the "Diamond" or the "Ruby" home model design and could have the home built on either a lake lot or a treed lot (with no water access).

In order to understand the data in Table 1.1, note that any data set provides information about some group of individual **elements,** which may be people, objects, events, or other entities. The information that a data set provides about its elements usually describes one or more characteristics of these elements.

> Any characteristic of an element is called a **variable.**

TABLE 1.1 **A Data Set Describing Five Home Sales** DS HomeSales

Home	Model Design	Lot Type	List Price	Selling Price
1	Diamond	Lake	$494,000	$494,000
2	Ruby	Treed	$447,000	$398,000
3	Diamond	Treed	$494,000	$440,000
4	Diamond	Treed	$494,000	$469,000
5	Ruby	Lake	$447,000	$447,000

For the data set in Table 1.1, each sold home is an element, and four variables are used to describe the homes. These variables are (1) the home model design, (2) the type of lot on which the home was built, (3) the list (asking) price, and (4) the (actual) selling price. Moreover, each home model design came with "everything included"—specifically, a complete, luxury interior package and a choice (at no price difference) of one of three different architectural exteriors. The builder made the list price of each home solely dependent on the model design. However, the builder gave various price reductions for homes built on treed lots.

The data in Table 1.1 are real (with some minor changes to protect privacy) and were provided by a business executive—a friend of the authors—who recently received a promotion and needed to move to central Florida. While searching for a new home, the executive and his family visited the luxury home community and decided they wanted to purchase a Diamond model on a treed lot. The list price of this home was $494,000, but the developer offered to sell it for an "incentive" price of $469,000. Intuitively, the incentive price's $25,000 savings off list price seemed like a good deal. However, the executive resisted making an immediate decision. Instead, he decided to collect data on the selling prices of new homes recently sold in the community and use the data to assess whether the developer might accept a lower offer. In order to collect "relevant data," the executive talked to local real estate professionals and learned that new homes sold in the community during the previous three months were a good indicator of current home value. Using real estate sales records, the executive also learned that five of the community's new homes had sold in the previous three months. The data given in Table 1.1 are the data that the executive collected about these five homes.

When the business executive examined Table 1.1, he noted that homes on lake lots had sold at their list price, but homes on treed lots had not. Because the executive and his family wished to purchase a Diamond model on a treed lot, the executive also noted that two Diamond models on treed lots had sold in the previous three months. One of these Diamond models had sold for the incentive price of $469,000, but the other had sold for a lower price of $440,000. Hoping to pay the lower price for his family's new home, the executive offered $440,000 for the Diamond model on the treed lot. Initially, the home builder turned down this offer, but two days later the builder called back and accepted the offer. The executive had used data to buy the new home for $54,000 less than the list price and $29,000 less than the incentive price!

LO1-2

Describe the difference between a quantitative variable and a qualitative variable.

Quantitative and qualitative variables

For any variable describing an element in a data set, we carry out a **measurement** to assign a value of the variable to the element. For example, in the real estate example, real estate sales records gave the actual selling price of each home to the nearest dollar. As another example, a credit card company might measure the time it takes for a cardholder's bill to be paid to the nearest day. Or, as a third example, an automaker might measure the gasoline mileage obtained by a car in city driving to the nearest one-tenth of a mile per gallon by conducting a mileage test on a driving course prescribed by the Environmental Protection Agency (EPA). If the possible values of a variable are numbers that represent quantities (that is, "how much" or "how many"), then the variable is said to be **quantitative.** For example, (1) the actual selling price of a home, (2) the payment time of a bill, (3) the gasoline mileage of a car, and (4) the 2016 payroll of a Major League Baseball team are all quantitative variables. Considering the last example, Table 1.2 gives the 2016 payroll (in millions of dollars) for each of the 30 Major League Baseball (MLB) teams. Moreover, Figure 1.1 portrays the team payrolls as a **dot plot.** In this plot, each team payroll is shown as a dot located on the real number

TABLE 1.2 2016 MLB Payrolls

DS MLB

Team	2016 Payroll
Los Angeles Dodgers	223
New York Yankees	213
Boston Red Sox	182
Detroit Tigers	172
San Francisco Giants	166
Washington Nationals	166
Los Angeles Angels	146
Texas Rangers	144
Philadelphia Phillies	133
Toronto Blue Jays	126
Seattle Mariners	123
St. Louis Cardinals	120
Cincinnati Reds	117
Chicago Cubs	117
Baltimore Orioles	116
Kansas City Royals	113
San Diego Padres	113
Minnesota Twins	108
New York Mets	100
Chicago White Sox	99
Milwaukee Brewers	99
Colorado Rockies	98
Atlanta Braves	88
Cleveland Indians	86
Pittsburgh Pirates	86
Miami Marlins	85
Oakland Athletics	80
Tampa Bay Rays	74
Arizona Diamondbacks	71
Houston Astros	69

Source: www.stevetheump.com, January 15, 2017.

FIGURE 1.1 A Dot Plot of 2016 MLB Payrolls (Payroll Is a Quantitative Variable)

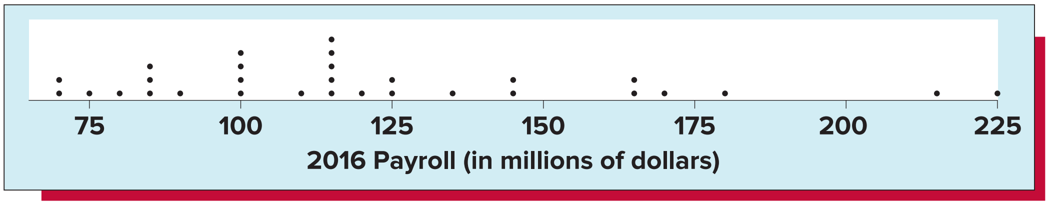


FIGURE 1.2 The Ten Most Popular Car Colors in the World for 2012 (Car Color Is a Qualitative Variable)

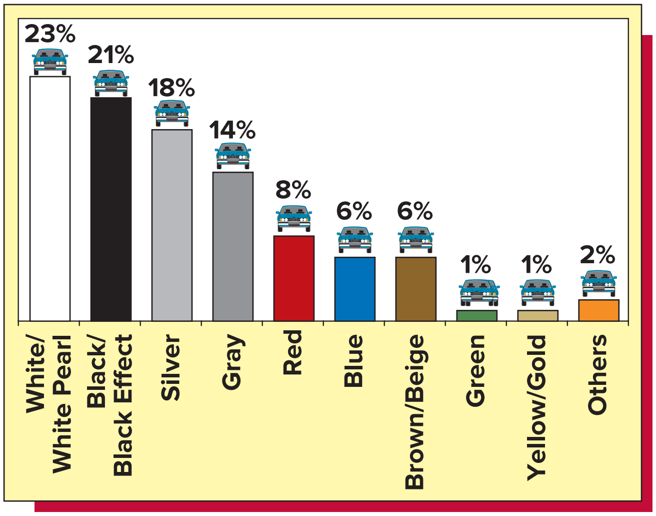
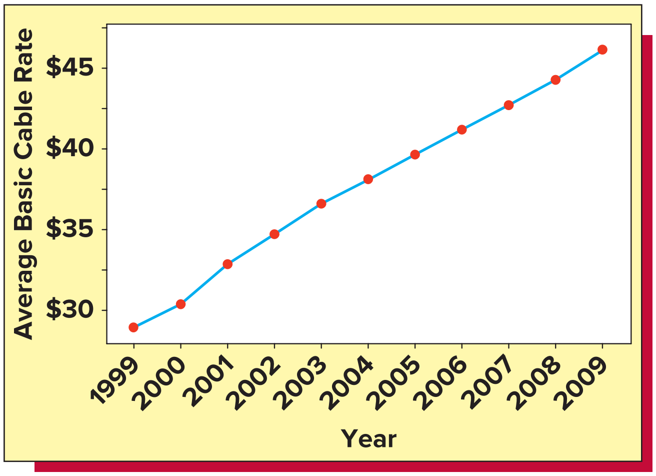


FIGURE 1.3 Time Series Plot of the Average Basic Cable Rates in the U.S. from 1999 to 2009 DS BasicCable



Source: Author created using data from http://autoweek.com/article/car-life/dupont-color-survey-puts-white-and-black-top-again (accessed September 12, 2013).

TABLE 1.3 The Average Basic Cable Rates in the U.S. from 1999 to 2009 DS BasicCable

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Cable Rate	$ 28.92	30.37	32.87	34.71	36.59	38.14	39.63	41.17	42.72	44.28	46.13

Source: U.S. Energy Information Administration, http://www.eia.gov/.

line—for example, the rightmost dot represents the payroll for the Los Angeles Dodgers. In general, the values of a quantitative variable are numbers on the real line. In contrast, if we simply record into which of several categories an element falls, then the variable is said to be **qualitative** or **categorical.** Examples of categorical variables include (1) a person's gender, (2) whether a person who purchases a product is satisfied with the product, (3) the type of lot on which a home is built, and (4) the color of a car.[1] Figure 1.2 illustrates the categories we might use for the qualitative variable "car color." This figure is a **bar chart** showing the 10 most popular (worldwide) car colors for 2012 and the percentages of cars having these colors.

Cross-sectional and time series data

Some statistical techniques are used to analyze *cross-sectional data,* while others are used to analyze *time series data.* **Cross-sectional data** are data collected at the same or approximately the same point in time. For example, suppose that a bank wishes to analyze last month's cell phone bills for its employees. Then, because the cell phone costs given by these bills are for different employees in the same month, the cell phone costs are cross-sectional data. **Time series data** are data collected over different time periods. For example, Table 1.3 presents the average basic cable television rate in the United States for each of the years 1999 to 2009. Figure 1.3 is a **time series plot**—also called a **runs plot**—of these data. Here we plot each cable rate on the vertical scale versus its corresponding time index (year) on the horizontal scale. For instance, the first cable rate ($28.92) is plotted versus 1999, the second cable rate ($30.37) is plotted versus 2000, and so forth. Examining the time series plot, we see that the cable rates increased substantially from 1999 to 2009. Finally, because the five homes in Table 1.1 were sold over a three-month period that represented a relatively stable real estate market, we can consider the data in Table 1.1 to essentially be cross-sectional data.

LO1-3
Describe the difference between cross-sectional data and time series data.

LO1-4
Construct and interpret a time series (runs) plot.

[1] Optional Section 1.6 discusses two types of quantitative variables (ratio and interval) and two types of qualitative variables (ordinal and nominative).

LO1-5

Identify the different types of data sources: existing data sources, experimental studies, and observational studies.

1.2 Data Sources, Data Warehousing, and Big Data

Primary data are data collected by an individual or business directly through planned **experimentation** or **observation. Secondary data** are data taken from an **existing source.**

Existing sources

Sometimes we can use data *already gathered* by public or private sources. The Internet is an obvious place to search for electronic versions of government publications, company reports, and business journals, but there is also a wealth of information available in the reference section of a good library or in county courthouse records.

If a business wishes to find demographic data about regions of the United States, a natural source is the U.S. Census Bureau's website at http://www.census.gov. Other useful websites for economic and financial data include the Federal Reserve at http://research.stlouisfed.org/fred2/ and the Bureau of Labor Statistics at http://stats.bls.gov/.

However, given the ease with which anyone can post documents, pictures, blogs, and videos on the Internet, not all sites are equally reliable. Some of the sources will be more useful, exhaustive, and error-free than others. Fortunately, search engines prioritize the lists and provide the most relevant and highly used sites first.

Obviously, performing such web searches costs next to nothing and takes relatively little time, but the tradeoff is that we are also limited in terms of the type of information we are able to find. Another option may be to use a private data source. Most companies keep and use employee records and information about their customers, products, processes (inventory, payroll, manufacturing, and accounting), and advertising results. If we have no affiliation with these companies, however, these data may be difficult to obtain.

Another alternative would be to contact a data collection agency, which typically incurs some kind of cost. You can either buy subscriptions or purchase individual company financial reports from agencies like Bloomberg and Dow Jones & Company. If you need to collect specific information, some companies, such as ACNielsen and Information Resources, Inc., can be hired to collect the information for a fee. Moreover, no matter what existing source you take data from, it is important to assess how reliable the data are by determing how, when, and where the data were collected.

©Pixtal/SuperStock

Experimental and observational studies

There are many instances when the data we need are not readily available from a public or private source. In cases like these, we need to collect the data ourselves. Suppose we work for a beverage company and want to assess consumer reactions to a new bottled water. Because the water has not been marketed yet, we may choose to conduct taste tests, focus groups, or some other market research. As another example, when projecting political election results, telephone surveys and exit polls are commonly used to obtain the information needed to predict voting trends. New drugs for fighting disease are tested by collecting data under carefully controlled and monitored experimental conditions. In many marketing, political, and medical situations of these sorts, companies sometimes hire outside consultants or statisticians to help them obtain appropriate data. Regardless of whether newly minted data are gathered in-house or by paid outsiders, this type of data collection requires much more time, effort, and expense than are needed when data can be found from public or private sources.

When initiating a study, we first define our variable of interest, or **response variable.** Other variables, typically called **factors,** that may be related to the response variable of interest will also be measured. When we are able to set or manipulate the values of these factors, we have an **experimental study.** For example, a pharmaceutical company might wish to determine the most appropriate daily dose of a cholesterol-lowering drug for patients having cholesterol levels that are too high. The company can perform an experiment in which one sample of patients receives a placebo; a second sample receives some low dose; a third a higher dose; and so forth. This is an experiment because the company controls the amount of drug each group receives. The optimal daily dose can be determined by analyzing the patients' responses to the different dosage levels given.

When analysts are unable to control the factors of interest, the study is **observational.** In studies of diet and cholesterol, patients' diets are not under the analyst's control. Patients are often unwilling or unable to follow prescribed diets; doctors might simply ask patients what they eat and then look for associations between the factor *diet* and the response variable *cholesterol level*.

Asking people what they eat is an example of performing a **survey.** In general, people in a survey are asked questions about their behaviors, opinions, beliefs, and other characteristics. For instance, shoppers at a mall might be asked to fill out a short questionnaire which seeks their opinions about a new bottled water. In other observational studies, we might simply observe the behavior of people. For example, we might observe the behavior of shoppers as they look at a store display, or we might observe the interactions between students and teachers.

Transactional data, data warehousing, and big data

LO1-6
Explain the basic ideas of data warehousing and big data.

With the increased use of online purchasing and with increased competition, businesses have become more aggressive about collecting information concerning customer transactions. Every time a customer makes an online purchase, more information is obtained than just the details of the purchase itself. For example, the web pages searched before making the purchase and the times that the customer spent looking at the different web pages are recorded. Similarly, when a customer makes an in-store purchase, store clerks often ask for the customer's address, zip code, e-mail address, and telephone number. By studying past customer behavior and pertinent demographic information, businesses hope to accurately predict customer response to different marketing approaches and leverage these predictions into increased revenues and profits. Dramatic advances in data capture, data transmission, and data storage capabilities are enabling organizations to integrate various databases into **data warehouses.** *Data warehousing* is defined as a process of centralized data management and retrieval and has as its ideal objective the creation and maintenance of a central repository for all of an organization's data. The huge capacity of data warehouses has given rise to the term **big data,** which refers to massive amounts of data, often collected at very fast rates in real time and in different forms and sometimes needing quick preliminary analysis for effective business decision making.

C EXAMPLE 1.1 The Disney Parks Case: Improving Visitor Experiences

Annually, approximately 100 million visitors spend time in Walt Disney parks around the world. These visitors could generate a lot of data, and in 2013, Walt Disney World Parks and Resorts introduced the wireless-tracking wristband *MagicBand* in Walt Disney World in Orlando, Florida.

©Bob Croslin

The MagicBands are linked to a credit card and serve as a park entry pass and hotel room key. They are part of the *McMagic*$^+$ system and wearing a band is completely voluntary. In addition to expediting sales transactions and hotel room access in the Disney theme parks, MagicBands provide visitors with easier access to FastPass lines for Disney rides and attractions. Each visitor to a Disney theme park may choose a FastPass for three rides or attractions per day. A FastPass allows a visitor to enter a line where there is virtually no waiting time. The McMagic$^+$ system automatically programs a visitor's FastPass selections into his or her MagicBand. As shown by the photo, a visitor simply places the MagicBand on his or her wrist next to a FastPass entry reader and is immediately admitted to the ride or attraction.

In return, the McMagic$^+$ system allows Disney to collect massive amounts of valuable data like real-time location, purchase history, riding patterns, and audience analysis and segmentation data. For example, the data tell Disney the types and ages of people who like specific attractions. To store, process, analyze and visualize all the data, Disney has constructed a gigantic data warehouse and a big data analysis platform. The data analysis allows Disney to improve daily park operations (by having the right numbers of staff on hand for the number of visitors currently in the park); to improve visitor experiences when choosing their "next" ride (by having large displays showing the waiting times for the park's rides); to improve its attraction offerings; and to tailor its marketing messages to different types of visitors.

Finally, although it collects massive amounts of data, Disney is very ethical in protecting the privacy of its visitors. First, as previously stated, visitors can choose not to wear a MagicBand. Moreover, visitors who do decide to wear one have control over the quantities of data collected, stored, and shared. Visitors can use a menu to specify whether Disney can send them personalized offers during or after their park visit. Parents also have to opt in before the characters in the park can address their children by name or use other personal information stored in the MagicBands.

Exercises for Sections 1.1 and 1.2

CONCEPTS connect

1.1 Define what we mean by a *variable,* and explain the difference between a quantitative variable and a qualitative (categorical) variable.

1.2 Below we list several variables. Which of these variables are quantitative and which are qualitative? Explain.

- **a** The dollar amount on an accounts receivable invoice.
- **b** The net profit for a company in 2017.
- **c** The stock exchange on which a company's stock is traded.
- **d** The national debt of the United States in 2017.
- **e** The advertising medium (radio, television, or print) used to promote a product.

1.3 **(1)** Discuss the difference between cross-sectional data and time series data. **(2)** If we record the total number of cars sold in 2017 by each of 10 car salespeople, are the data cross-sectional or time series data? **(3)** If we record the total number of cars sold by a particular car salesperson in each of the years 2013, 2014, 2015, 2016, and 2017, are the data cross-sectional or time series data?

1.4 Consider a medical study that is being performed to test the effect of smoking on lung cancer. Two groups of subjects are identified; one group has lung cancer and the other one doesn't. Both are asked to fill out a questionnaire containing questions about their age, sex, occupation, and number of cigarettes smoked per day. **(1)** What is the response variable? **(2)** Which are the factors? **(3)** What type of study is this (experimental or observational)?

1.5 What is a data warehouse? What does the term *big data* mean?

METHODS AND APPLICATIONS

1.6 Consider the five homes in Table 1.1. What do you think you would have to pay for a Ruby model on a treed lot?

1.7 Consider the five homes in Table 1.1. What do you think you would have to pay for a Diamond model on a lake lot? For a Ruby model on a lake lot?

1.8 The number of Bismark X-12 electronic calculators sold at Smith's Department Stores over the past 24 months have been: 197, 211, 203, 247, 239, 269, 308, 262, 258, 256, 261, 288, 296, 276, 305, 308, 356, 393, 363, 386, 443, 308, 358, and 384. Make a time series plot of these data. That is, plot 197 versus month 1, 211 versus month 2, and so forth. What does the time series plot tell you? DS CalcSale

1.3 Populations, Samples, and Traditional Statistics

LO1-7
Describe the difference between a population and a sample.

We often collect data in order to study a population.

> A **population** is the set of all elements about which we wish to draw conclusions.

Examples of populations include (1) all of last year's graduates of Dartmouth College's Master of Business Administration program, (2) all current MasterCard cardholders, and (3) all Buick LaCrosses that have been or will be produced this year.

We usually focus on studying one or more variables describing the population elements. If we carry out a measurement to assign a value of a variable to each and every population element, we have a *population of measurements* (sometimes called *observations*). If the population is small, it is reasonable to do this. For instance, if 150 students graduated last year from the Dartmouth College MBA program, it might be feasible to survey the graduates and to record all of their starting salaries. In general:

> If we examine all of the population measurements, we say that we are conducting a **census** of the population.

Often the population that we wish to study is very large, and it is too time-consuming or costly to conduct a census. In such a situation, we select and analyze a subset (or portion) of the population elements.

> A **sample** is a subset of the elements of a population.

For example, suppose that 8,742 students graduated last year from a large state university. It would probably be too time-consuming to take a census of the population of all of their starting salaries. Therefore, we would select a sample of graduates, and we would obtain and record their starting salaries. When we measure a characteristic of the elements in a sample, we have a **sample of measurements.**

We often wish to describe a population or sample.

LO1-8
Distinguish between descriptive statistics and statistical inference.

> **Descriptive statistics** is the science of describing the important aspects of a set of measurements.

As an example, if we are studying a set of starting salaries, we might wish to describe (1) what a typical salary might be and (2) how much the salaries vary from each other.

When the population of interest is small and we can conduct a census of the population, we will be able to directly describe the important aspects of the population measurements. However, if the population is large and we need to select a sample from it, then we use what we call **statistical inference.**

> **Statistical inference** is the science of using a sample of measurements to make generalizations about the important aspects of a population of measurements.

For instance, we might use the starting salaries recorded for a sample of the 8,742 students who graduated last year from a large state university to *estimate* the typical starting salary and the variation of the starting salaries for the entire population of the 8,742 graduates. Or General Motors might use a sample of Buick LaCrosses produced this year to estimate the typical EPA combined city and highway driving mileage and the variation of these mileages for all LaCrosses that have been or will be produced this year.

What we might call **traditional statistics** consists of a set of concepts and techniques that are used to describe populations and samples and to make statistical inferences about populations by using samples. Much of this book is devoted to traditional statistics, and in the next section we will discuss **random sampling** (or approximately random sampling). However, traditional statistics is sometimes not sufficient to analyze big data, which (we recall) refers to massive amounts of data often collected at very fast rates in real time and sometimes needing quick preliminary analysis for effective business decision making. For this reason, two related extensions of traditional statistics—**business analytics** and **data mining**—have been developed to help analyze big data. In optional Section 1.5 we will begin to discuss business analytics and data mining. As one example of using business analytics, we will see how Disney uses the large amount of data it collects every day concerning the riding patterns of its visitors. These data are used to keep its visitors informed of the current waiting times for different rides, which helps patrons select the next ride to go on or attraction to attend.

1.4 Random Sampling and Three Case Studies That Illustrate Statistical Inference

LO1-9
Explain the concept of random sampling and select a random sample.

Random sampling

If the information contained in a sample is to accurately reflect the population under study, the sample should be **randomly selected** from the population. To intuitively illustrate random sampling, suppose that a small company employs 15 people and wishes to randomly select two of them to attend a convention. To make the random selections, we number the employees from 1 to 15, and we place in a hat 15 identical slips of paper numbered from 1 to 15. We thoroughly mix the slips of paper in the hat and, blindfolded, choose one. The number on the chosen slip of paper identifies the first randomly selected employee. Then, still blindfolded, we choose another slip of paper from the hat. The number on the second slip identifies the second randomly selected employee.

Of course, when the population is large, it is not practical to randomly select slips of paper from a hat. For instance, experience has shown that thoroughly mixing slips of paper (or the like) can be difficult. Further, dealing with many identical slips of paper would be cumbersome and time-consuming. For these reasons, statisticians have developed more efficient and accurate methods for selecting a random sample. To discuss these methods we let n denote the number of elements in a sample. We call n the **sample size.** We now define a random sample of n elements and explain how to select such a sample.[2]

1 If we select n elements from a population in such a way that every set of n elements in the population has the same chance of being selected, then the n elements we select are said to be a **random sample.**

2 In order to select a random sample of n elements from a population, we make n *random selections*—one at a time—from the population. On each **random selection,** we give every element remaining in the population for that selection the same chance of being chosen.

In making random selections from a population, we can sample *with or without replacement.* If we **sample with replacement,** we place the element chosen on any particular selection back into the population. Thus, we give this element a chance to be chosen on any succeeding selection. If we **sample without replacement,** we do not place the element chosen on a particular selection back into the population. Thus, we do not give this element a chance to be chosen on any succeeding selection. **It is best to sample without replacement.** Intuitively, this is because choosing the sample without replacement guarantees that all of the elements in the sample will be different, and thus we will have the fullest possible look at the population.

We now introduce three case studies that illustrate (1) the need for a random (or approximately random) sample, (2) how to select the needed sample, and (3) the use of the sample in making statistical inferences.

C EXAMPLE 1.2 The Cell Phone Case: Reducing Cellular Phone Costs

Part 1: The Cost of Company Cell Phone Use Rising cell phone costs have forced companies having large numbers of cellular users to hire services to manage their cellular and other wireless resources. These cellular management services use sophisticated software and mathematical models to choose cost-efficient cell phone plans for their clients. One such firm, mindWireless of Austin, Texas, specializes in automated wireless cost management.

[2] Actually, there are several different kinds of random samples. The type we will define is sometimes called a *simple random sample*. For brevity's sake, however, we will use the term *random sample*.

According to Kevin Whitehurst, co-founder of mindWireless, cell phone carriers count on *overage*—using more minutes than one's plan allows—and *underage*—using fewer minutes than those already paid for—to deliver almost half of their revenues.[3] As a result, a company's typical cost of cell phone use can be excessive—18 cents per minute or more. However, Mr. Whitehurst explains that by using mindWireless automated cost management to select calling plans, this cost can be reduced to 12 cents per minute or less.

In this case we consider a bank that wishes to decide whether to hire a cellular management service to choose its employees' calling plans. While the bank has over 10,000 employees on many different types of calling plans, a cellular management service suggests that by studying the calling patterns of cellular users on 500-minute-per-month plans, the bank can accurately assess whether its cell phone costs can be substantially reduced. The bank has 2,136 employees on a variety of 500-minute-per-month plans with different basic monthly rates, different overage charges, and different additional charges for long distance and roaming. It would be extremely time-consuming to analyze in detail the cell phone bills of all 2,136 employees. Therefore, the bank will estimate its cellular costs for the 500-minute plans by analyzing last month's cell phone bills for a *random sample* of 100 employees on these plans.[4]

Part 2: Selecting a Random Sample The first step in selecting a random sample is to obtain a numbered list of the population elements. This list is called a **frame.** Then we can use a *random number table* or *computer-generated random numbers* to make random selections from the numbered list. Therefore, in order to select a random sample of 100 employees from the population of 2,136 employees on 500-minute-per-month cell phone plans, the bank will make a numbered list of the 2,136 employees on 500-minute plans. The bank can then use a **random number table,** such as Table 1.4(a), to select the random sample. To see how this is done, note that any single-digit number in the table has been chosen in such a way that any of the single-digit numbers between 0 and 9 had the same chance of being chosen. For this reason, we say that any single-digit number in the table is a **random number** between 0 and 9. Similarly, any two-digit number in the table is a random number between 00 and 99, any three-digit number in the table is a random number between 000 and 999, and so forth. Note that the table entries are segmented into groups of five to make the table easier to read. Because the total number of employees on 500-minute cell phone plans (2,136) is a four-digit number, we arbitrarily select any set of four digits in the table (we have circled these digits). This number, which is 0511, identifies the first randomly selected employee. Then, moving in any direction from the 0511 (up, down, right, or left—it does not matter which), we select additional sets of four digits. These succeeding sets of digits identify additional randomly selected employees. Here we arbitrarily move down from 0511 in the table. The first seven sets of four digits we obtain are

0511 7156 0285 4461 3990 4919 1915

(See Table 1.4(a)—these numbers are enclosed in a rectangle.) Because there are no employees numbered 7156, 4461, 3990, or 4919 (remember only 2,136 employees are on 500-minute plans), we ignore these numbers. This implies that the first three randomly selected employees are those numbered 0511, 0285, and 1915. Continuing this procedure, we can obtain the entire random sample of 100 employees. Notice that, because we are sampling without replacement, we should ignore any set of four digits previously selected from the random number table.

While using a random number table is one way to select a random sample, this approach has a disadvantage that is illustrated by the current situation. Specifically, because most four-digit random numbers are not between 0001 and 2136, obtaining 100 different, four-digit random numbers between 0001 and 2136 will require ignoring a large number of random numbers in the random number table, and we will in fact need to use a random number table that is larger than Table 1.4(a). Although larger random number tables are readily available in books of mathematical and statistical tables, a good alternative is to use a computer

[3] The authors would like to thank Kevin Whitehurst for help in developing this case.

[4] In Chapter 9 we will discuss how to plan the *sample size*—the number of elements (for example, 100) that should be included in a sample. Throughout this book we will take large enough samples to allow us to make reasonably accurate statistical inferences.

TABLE 1.4 Random Numbers

(a) A portion of a random number table

33276	85590	79936	56865	05859	90106	78188
03427	90511	69445	18663	72695	52180	90322
92737	27156	33488	36320	17617	30015	74952
85689	20285	52267	67689	93394	01511	89868
08178	74461	13916	47564	81056	97735	90707
51259	63990	16308	60756	92144	49442	40719
60268	44919	19885	55322	44819	01188	55157
94904	01915	04146	18594	29852	71585	64951
58586	17752	14513	83149	98736	23495	35749
09998	19509	06691	76988	13602	51851	58104
14346	61666	30168	90229	04734	59193	32812
74103	15227	25306	76468	26384	58151	44592
24200	64161	38005	94342	28728	35806	22851
87308	07684	00256	45834	15398	46557	18510
07351	86679	92420	60952	61280	50001	94953

(b) Minitab output of 100 different, four-digit random numbers between 1 and 2136

705	1131	169	1703	1709	609
1990	766	1286	1977	222	43
1007	1902	1209	2091	1742	1152
111	69	2049	1448	659	338
1732	1650	7	388	613	1477
838	272	1227	154	18	320
1053	1466	2087	265	2107	1992
582	1787	2098	1581	397	1099
757	1699	567	1255	1959	407
354	1567	1533	1097	1299	277
663	40	585	1486	1021	532
1629	182	372	1144	1569	1981
1332	1500	743	1262	1759	955
1832	378	728	1102	667	1885
514	1128	1046	116	1160	1333
831	2036	918	1535	660	
928	1257	1468	503	468	

software package, which can generate random numbers that are between whatever values we specify. For example, Table 1.4(b) gives the Minitab output of 100 different, four-digit random numbers that are between 0001 and 2136 (note that the "leading 0's" are not included in these four-digit numbers). If used, the random numbers in Table 1.4(b) would identify the 100 employees that form the random sample. For example, the first three randomly selected employees would be employees 705, 1990, and 1007.

Finally, note that computer software packages sometimes generate the same random number twice and thus are sampling with replacement. Because we wished to randomly select 100 employees without replacement, we had Minitab generate more than 100 (actually, 110) random numbers. We then ignored the repeated random numbers to obtain the 100 different random numbers in Table 1.4(b).

Part 3: A Random Sample and Inference When the random sample of 100 employees is chosen, the number of cellular minutes used by each sampled employee during last month (the employee's *cellular usage*) is found and recorded. The 100 cellular-usage figures are given in Table 1.5. Looking at this table, we can see that there is substantial overage and underage—many employees used far more than 500 minutes, while many others failed to use all of the 500 minutes allowed by their plan. In Chapter 3 we will use these 100 usage figures to estimate the bank's cellular costs and decide whether the bank should hire a cellular management service.

TABLE 1.5 A Sample of Cellular Usages (in Minutes) for 100 Randomly Selected Employees
DS CellUse

75	485	37	547	753	93	897	694	797	477
654	578	504	670	490	225	509	247	597	173
496	553	0	198	507	157	672	296	774	479
0	822	705	814	20	513	546	801	721	273
879	433	420	521	648	41	528	359	367	948
511	704	535	585	341	530	216	512	491	0
542	562	49	505	461	496	241	624	885	259
571	338	503	529	737	444	372	555	290	830
719	120	468	730	853	18	479	144	24	513
482	683	212	418	399	376	323	173	669	611

EXAMPLE 1.3 The Marketing Research Case: Rating a Bottle Design

Part 1: Rating a Bottle Design The design of a package or bottle can have an important effect on a company's bottom line. In this case a brand group wishes to research consumer reaction to a new bottle design for a popular soft drink. Because it is impossible to show the new bottle design to "all consumers," the brand group will use the *mall intercept method* to select a sample of 60 consumers. On a particular Saturday, the brand group will choose a shopping mall and a sampling time so that shoppers at the mall during the sampling time are a representative cross-section of all consumers. Then, shoppers will be intercepted as they walk past a designated location, will be shown the new bottle, and will be asked to rate the bottle image. For each consumer interviewed, a bottle image **composite score** will be found by adding the consumer's numerical responses to the five questions shown in Figure 1.4. It follows that the minimum possible bottle image composite score is 5 (resulting from a response of 1 on all five questions) and the maximum possible bottle image composite score is 35 (resulting from a response of 7 on all five questions). Furthermore, experience has shown that the smallest acceptable bottle image composite score for a successful bottle design is 25.

Part 2: Selecting an Approximately Random Sample Because it is not possible to list and number all of the shoppers who will be at the mall on this Saturday, we cannot select a random sample of these shoppers. However, we can select an *approximately* random sample of these shoppers. To see one way to do this, note that there are 6 ten-minute intervals during each hour, and thus there are 60 ten-minute intervals during the 10-hour period from 10 A.M. to 8 P.M.—the time when the shopping mall is open. Therefore, one way to select an approximately random sample is to choose a particular location at the mall that most shoppers will walk by and then randomly select—at the beginning of each ten-minute period—one of the first shoppers who walks by the location. Here, although we could randomly select one person from any reasonable number of shoppers who walk by, we will (arbitrarily) randomly select one of the first five shoppers who walk by. For example, starting in the upper left-hand corner of Table 1.4(a) and proceeding down the first column, note that the first three random numbers between 1 and 5 are 3, 5, and 1. This implies that (1) at 10 A.M. we would select the 3rd customer who walks by; (2) at 10:10 A.M. we would select the 5th shopper who walks by; (3) at 10:20 A.M. we would select the 1st customer who walks by, and so forth. Furthermore, assume that the composite score ratings of the new bottle design that would be given by all shoppers at the mall on the Saturday are representative of the composite score ratings that would be given by all possible consumers. It then follows that the composite score ratings given by the 60 sampled shoppers can be regarded as an approximately random sample that can be used to make statistical inferences about the population of all possible consumer composite score ratings.

Part 3: The Approximately Random Sample and Inference When the brand group uses the mall intercept method to interview a sample of 60 shoppers at a mall on a particular Saturday, the 60 bottle image composite scores in Table 1.6 are obtained. Because these scores

FIGURE 1.4 The Bottle Design Survey Instrument

Please circle the response that most accurately describes whether you agree or disagree with each statement about the bottle you have examined.

Statement	Strongly Disagree						Strongly Agree
The size of this bottle is convenient.	1	2	3	4	5	6	7
The contoured shape of this bottle is easy to handle.	1	2	3	4	5	6	7
The label on this bottle is easy to read.	1	2	3	4	5	6	7
This bottle is easy to open.	1	2	3	4	5	6	7
Based on its overall appeal, I like this bottle design.	1	2	3	4	5	6	7

TABLE 1.6 A Sample of Bottle Design Ratings (Composite Scores for a Sample of 60 Shoppers)
DS Design

34	33	33	29	26	33	28	25	32	33
32	25	27	33	22	27	32	33	32	29
24	30	20	34	31	32	30	35	33	31
32	28	30	31	31	33	29	27	34	31
31	28	33	31	32	28	26	29	32	34
32	30	34	32	30	30	32	31	29	33

vary from a minimum of 20 to a maximum of 35, we might infer that *most* consumers would rate the new bottle design between 20 and 35. Furthermore, 57 of the 60 composite scores are at least 25. Therefore, we might estimate that a proportion of 57/60 = .95 (that is, 95 percent) of all consumers would give the bottle design a composite score of at least 25. In future chapters we will further analyze the composite scores.

Processes

Sometimes we are interested in studying the population of all of the elements that will be or could potentially be produced by a *process*.

> A **process** is a sequence of operations that takes inputs (labor, materials, methods, machines, and so on) and turns them into outputs (products, services, and the like).

Processes produce output *over time*. For example, this year's Buick LaCrosse manufacturing process produces LaCrosses over time. Early in the model year, General Motors might wish to study the population of the city driving mileages of all Buick LaCrosses that will be produced during the model year. Or, even more hypothetically, General Motors might wish to study the population of the city driving mileages of all LaCrosses that could *potentially* be produced by this model year's manufacturing process. The first population is called a **finite population** because only a finite number of cars will be produced during the year. The second population is called an **infinite population** because the manufacturing process that produces this year's model could in theory always be used to build "one more car." That is, theoretically there is no limit to the number of cars that could be produced by this year's process. There are a multitude of other examples of finite or infinite hypothetical populations. For instance, we might study the population of all waiting times that will or could potentially be experienced by patients of a hospital emergency room. Or we might study the population of all the amounts of grape jelly that will be or could potentially be dispensed into 16-ounce jars by an automated filling machine. To study a population of potential process observations, we sample the process—often at equally spaced time points—over time.

C EXAMPLE 1.4 The Car Mileage Case: Estimating Mileage

Part 1: Auto Fuel Economy Personal budgets, national energy security, and the global environment are all affected by our gasoline consumption. Hybrid and electric cars are a vital part of a long-term strategy to reduce our nation's gasoline consumption. However, until use of these cars is more widespread and affordable, the most effective way to conserve gasoline is to design gasoline-powered cars that are more fuel efficient.[5] In the short term, "that will give you the biggest bang for your buck," says David Friedman, research director of the Union of Concerned Scientists' Clean Vehicle Program.[6]

In this case study we consider a tax credit offered by the federal government to automakers for improving the fuel economy of gasoline-powered midsize cars. According to *The Fuel Economy Guide—2017 Model Year,* virtually every gasoline-powered midsize car equipped with an automatic transmission and a six-cylinder engine has an EPA combined city and

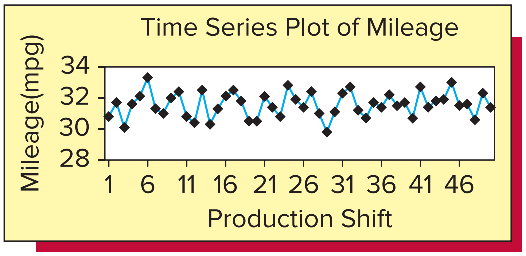
[5,6] Bryan Walsh, "Plugged In," *Time*, September 29, 2008 (see page 56).

TABLE 1.7 A Sample of 50 Mileages DS GasMiles

30.8	30.8	32.1	32.3	32.7
31.7	30.4	31.4	32.7	31.4
30.1	32.5	30.8	31.2	31.8
31.6	30.3	32.8	30.7	31.9
32.1	31.3	31.9	31.7	33.0
33.3	32.1	31.4	31.4	31.5
31.3	32.5	32.4	32.2	31.6
31.0	31.8	31.0	31.5	30.6
32.0	30.5	29.8	31.7	32.3
32.4	30.5	31.1	30.7	31.4

Note: Time order is given by reading down the columns from left to right.

FIGURE 1.5 A Time Series Plot of the 50 Mileages



highway mileage estimate of 26 miles per gallon (mpg) or less.[7] As a matter of fact, when this book was written in 2017, the mileage leader in this category was the Nissan Altima, which registered a combined city and highway mileage of 26 mpg. While fuel economy has seen improvement in almost all car categories, the EPA has concluded that an additional 5 mpg increase in fuel economy is significant and feasible.[8] Therefore, suppose that the government has decided to offer the tax credit to any automaker selling a midsize model with an automatic transmission and a six-cylinder engine that achieves an EPA combined city and highway mileage estimate of at least 31 mpg.

Part 2: Sampling a Process Consider an automaker that has recently introduced a new midsize model with an automatic transmission and a six-cylinder engine and wishes to demonstrate that this new model qualifies for the tax credit. In order to study the population of all cars of this type that will or could potentially be produced, the automaker will choose a sample of 50 of these cars. The manufacturer's production operation runs 8-hour shifts, with 100 midsize cars produced on each shift. When the production process has been fine-tuned and all start-up problems have been identified and corrected, the automaker will select one car at random from each of 50 consecutive production shifts. Once selected, each car is to be subjected to an EPA test that determines the EPA combined city and highway mileage of the car.

To randomly select a car from a particular production shift, we number the 100 cars produced on the shift from 00 to 99 and use a random number table or a computer software package to obtain a random number between 00 and 99. For example, starting in the upper left-hand corner of Table 1.4(a) and proceeding down the two leftmost columns, we see that the first three random numbers between 00 and 99 are 33, 3, and 92. This implies that we would select car 33 from the first production shift, car 3 from the second production shift, car 92 from the third production shift, and so forth. Moreover, because a new group of 100 cars is produced on each production shift, repeated random numbers would not be discarded. For example, if the 15th and 29th random numbers are both 7, we would select the 7th car from the 15th production shift and the 7th car from the 29th production shift.

Part 3: The Sample and Inference Suppose that when the 50 cars are selected and tested, the sample of 50 EPA combined mileages shown in Table 1.7 is obtained. A time series plot of the mileages is given in Figure 1.5. Examining this plot, we see that, although the mileages vary over time, they do not seem to vary in any unusual way. For example, the mileages do not tend to either decrease or increase (as did the basic cable rates in Figure 1.3) over time. This intuitively verifies that the midsize car manufacturing process is producing consistent car mileages over time, and thus we can regard the 50 mileages as an approximately random sample that can be used to make statistical inferences about the population of all

[7] The "26 miles per gallon (mpg) or less" figure relates to midsize cars with an automatic transmission *and* at least a six-cylinder, 3.5-liter engine. Therefore, when we refer to a midsize car with an automatic transmission in future discussions, we are assuming that the midsize car also has at least a six-cylinder, 3.5-liter engine.

[8] The authors wish to thank Jeff Alson of the EPA for this information.

possible midsize car mileages.[9] Therefore, because the 50 mileages vary from a minimum of 29.8 mpg to a maximum of 33.3 mpg, we might conclude that most midsize cars produced by the manufacturing process will obtain between 29.8 mpg and 33.3 mpg. Moreover, because 38 out of the 50 mileages—or 76 percent of the mileages—are greater than or equal to the tax credit standard of 31 mpg, we have some evidence that the "typical car" produced by the process will meet or exceed the tax credit standard. We will further evaluate this evidence in later chapters.

Probability sampling

Random (or approximately random) sampling—as well as the more advanced kinds of sampling discussed in optional Section 1.7—are types of *probability sampling.* In general, **probability sampling** is sampling where we know the chance (or probability) that each element in the population will be included in the sample. If we employ probability sampling, the sample obtained can be used to make valid statistical inferences about the sampled population. However, if we do not employ probability sampling, we cannot make valid statistical inferences.

One type of sampling that is not probability sampling is **convenience sampling,** where we select elements because they are easy or convenient to sample. For example, if we select people to interview because they look "nice" or "pleasant," we are using convenience sampling. Another example of convenience sampling is the use of **voluntary response samples,** which are frequently employed by television and radio stations and newspaper columnists. In such samples, participants self-select—that is, whoever wishes to participate does so (usually expressing some opinion). These samples overrepresent people with strong (usually negative) opinions. For example, the advice columnist Ann Landers once asked her readers, "If you had it to do over again, would you have children?" Of the nearly 10,000 parents who *voluntarily* responded, 70 percent said that they would not. A probability sample taken a few months later found that 91 percent of parents would have children again.

Another type of sampling that is not probability sampling is **judgment sampling,** where a person who is extremely knowledgeable about the population under consideration selects population elements that he or she feels are most representative of the population. Because the quality of the sample depends upon the judgment of the person selecting the sample, it is dangerous to use the sample to make statistical inferences about the population.

To conclude this section, we consider a classic example where two types of sampling errors doomed a sample's ability to make valid statistical inferences. This example occurred prior to the presidential election of 1936, when the *Literary Digest* predicted that Alf Landon would defeat Franklin D. Roosevelt by a margin of 57 percent to 43 percent. Instead, Roosevelt won the election in a landslide. *Literary Digest*'s first error was to send out sample ballots (actually, 10 million ballots) to people who were mainly selected from the *Digest*'s subscription list and from telephone directories. In 1936 the country had not yet recovered from the Great Depression, and many unemployed and low-income people did not have phones or subscribe to the *Digest*. The *Digest*'s sampling procedure excluded these people, who overwhelmingly voted for Roosevelt. Second, only 2.3 million ballots were returned, resulting in the sample being a voluntary response survey. At the same time, George Gallup, founder of the Gallup Poll, was beginning to establish his survey business. He used a probability sample to correctly predict Roosevelt's victory. In optional Section 1.8 we discuss various issues related to designing surveys and more about the errors that can occur in survey samples.

Ethical guidelines for statistical practice

The American Statistical Association, the leading U.S. professional statistical association, has developed the report "Ethical Guidelines for Statistical Practice."[10] This report provides information that helps statistical practitioners to consistently use ethical statistical practices

[9] In Chapter 20 (on the website) we will discuss more precisely how to assess whether a process is operating consistently over time.

[10] American Statistical Association, "Ethical Guidelines for Statistical Practice," 1999.

and that helps users of statistical information avoid being misled by unethical statistical practices. Unethical statistical practices can take a variety of forms, including:

- **Improper sampling** Purposely selecting a biased sample—for example, using a nonrandom sampling procedure that overrepresents population elements supporting a desired conclusion or that underrepresents population elements not supporting the desired conclusion—is unethical. In addition, discarding already sampled population elements that do not support the desired conclusion is unethical. More will be said about proper and improper sampling later in this chapter.
- **Misleading charts, graphs, and descriptive measures** In Section 2.7, we will present an example of how misleading charts and graphs can distort the perception of changes in salaries over time. Using misleading charts or graphs to make the salary changes seem much larger or much smaller than they really are is unethical. In Section 3.1, we will present an example illustrating that many populations of individual or household incomes contain a small percentage of very high incomes. These very high incomes make the *population mean income* substantially larger than the *population median income*. In this situation we will see that the population median income is a better measure of the typical income in the population. Using the population mean income to give an inflated perception of the typical income in the population is unethical.
- **Inappropriate statistical analysis or inappropriate interpretation of statistical results** The American Statistical Association report emphasizes that selecting many different samples and running many different tests can eventually (by random chance alone) produce a result that makes a desired conclusion seem to be true, when the conclusion really isn't true. Therefore, continuing to sample and run tests until a desired conclusion is obtained and not reporting previously obtained results that do not support the desired conclusion is unethical. Furthermore, we should always report our sampling procedure and sample size and give an estimate of the reliability of our statistical results. Estimating this reliability will be discussed in Chapter 8 and beyond.

The above examples are just an introduction to the important topic of unethical statistical practices. The American Statistical Association report contains 67 guidelines organized into eight areas involving general professionalism and ethical responsibilities. These include responsibilities to clients, to research team colleagues, to research subjects, and to other statisticians, as well as responsibilities in publications and testimony and responsibilities of those who employ statistical practitioners.

Exercises for Sections 1.3 and 1.4

CONCEPTS

1.9 **(1)** Define a *population*. **(2)** Give an example of a population that you might study when you start your career after graduating from college. **(3)** Explain the difference between a census and a sample.

1.10 Explain each of the following terms:

a Descriptive statistics.
b Statistical inference.
c Random sample.
d Process.

1.11 Explain why sampling without replacement is preferred to sampling with replacement.

METHODS AND APPLICATIONS

1.12 In the page margin, we list 15 companies that have historically performed well in the food, drink, and tobacco industries. Consider the random numbers given in the random number table of Table 1.4(a). Starting in the upper left corner of Table 1.4(a) and moving down the two leftmost columns, we see that the first three two-digit numbers obtained are: 33, 03, and 92. Starting with these three random numbers, and moving down the two leftmost columns of Table 1.4(a) to find more two-digit random numbers, use Table 1.4(a) to randomly select five of these companies to be interviewed in detail about their business strategies. Hint: Note that we have numbered the companies from 1 to 15.

Companies:

1	Altria Group
2	PepsiCo
3	Coca-Cola
4	Archer Daniels
5	Anheuser-Bush
6	General Mills
7	Sara Lee
8	Coca-Cola Enterprises
9	Reynolds American
10	Kellogg
11	ConAgra Foods
12	HJ Heinz
13	Campbell Soup
14	Pepsi Bottling Group
15	Tyson Foods

FIGURE 1.6 The Video Game Satisfaction Survey Instrument

Statement	Strongly Disagree 1	2	3	4	5	6	Strongly Agree 7
The game console of the XYZ-Box is well designed.	1	2	3	4	5	6	7
The game controller of the XYZ-Box is easy to handle.	1	2	3	4	5	6	7
The XYZ-Box has high-quality graphics capabilities.	1	2	3	4	5	6	7
The XYZ-Box has high-quality audio capabilities.	1	2	3	4	5	6	7
The XYZ-Box serves as a complete entertainment center.	1	2	3	4	5	6	7
There is a large selection of XYZ-Box games to choose from.	1	2	3	4	5	6	7
I am totally satisfied with my XYZ-Box game system.	1	2	3	4	5	6	7

1.13 THE VIDEO GAME SATISFACTION RATING CASE DS VideoGame

A company that produces and markets video game systems wishes to assess its customers' level of satisfaction with a relatively new model, the XYZ-Box. In the six months since the introduction of the model, the company has received 73,219 warranty registrations from purchasers. The company will randomly select 65 of these registrations and will conduct telephone interviews with the purchasers. Specifically, each purchaser will be asked to state his or her level of agreement with each of the seven statements listed on the survey instrument given in Figure 1.6. Here, the level of agreement for each statement is measured on a 7-point Likert scale. Purchaser satisfaction will be measured by adding the purchaser's responses to the seven statements. It follows that for each consumer the minimum composite score possible is 7 and the maximum is 49. Furthermore, experience has shown that a purchaser of a video game system is "very satisfied" if his or her composite score is at least 42.

a Assume that the warranty registrations are numbered from 1 to 73,219 in a computer. Starting in the upper left corner of Table 1.4(a) and moving down the five leftmost columns, we see that the first three five-digit numbers obtained are 33276, 03427, and 92737. Starting with these three random numbers and moving down the five leftmost columns of Table 1.4(a) to find more five-digit random numbers, use Table 1.4(a) to randomly select the numbers of the first 10 warranty registrations to be included in the sample of 65 registrations.

b Suppose that when the 65 customers are interviewed, their composite scores are as given in Table 1.8. Using the largest and smallest observations in the data, estimate limits between which most of the 73,219 composite scores would fall. Also, estimate the proportion of the 73,219 composite scores that would be at least 42.

TABLE 1.8 Composite Scores for the Video Game Satisfaction Rating Case DS VideoGame

39	44	46	44	44
45	42	45	44	42
38	46	45	45	47
42	40	46	44	43
42	47	43	46	45
41	44	47	48	
38	43	43	44	
42	45	41	41	
46	45	40	45	
44	40	43	44	
40	46	44	44	
39	41	41	44	
40	43	38	46	
42	39	43	39	
45	43	36	41	

1.14 THE BANK CUSTOMER WAITING TIME CASE DS WaitTime

A bank manager has developed a new system to reduce the time customers spend waiting to be served by tellers during peak business hours. Typical waiting times during peak business hours under the current system are roughly 9 to 10 minutes. The bank manager hopes that the new system will lower typical waiting times to less than 6 minutes and wishes to evaluate the new system. When the new system is operating consistently over time, the bank manager decides to select a sample of 100 customers that need teller service during peak business hours. Specifically, for each of 100 peak business hours, the first customer that starts waiting for teller service at or after a randomly selected time during the hour will be chosen.

a Consider the peak business hours from 2:00 P.M. to 2:59 P.M., from 3:00 P.M. to 3:59 P.M., from 4:00 P.M. to 4:59 P.M., and from 5:00 P.M. to 5:59 P.M. on a particular day. Also, assume that a computer software system generates the following four random numbers between 00 and 59: 32, 00, 18, and 47. This implies that the randomly selected times during the first three peak business hours are 2:32 P.M., 3:00 P.M., and 4:18 P.M. What is the randomly selected time during the fourth peak business hour?

b When each customer is chosen, the number of minutes the customer spends waiting for teller service is recorded. The 100 waiting times that are observed are given in Table 1.9. Using the largest and smallest observations in the data,

TABLE 1.9 Waiting Times (in Minutes) for the Bank Customer Waiting Time Case DS WaitTime

1.6	6.2	3.2	5.6	7.9	6.1	7.2
6.6	5.4	6.5	4.4	1.1	3.8	7.3
5.6	4.9	2.3	4.5	7.2	10.7	4.1
5.1	5.4	8.7	6.7	2.9	7.5	6.7
3.9	.8	4.7	8.1	9.1	7.0	3.5
4.6	2.5	3.6	4.3	7.7	5.3	6.3
6.5	8.3	2.7	2.2	4.0	4.5	4.3
6.4	6.1	3.7	5.8	1.4	4.5	3.8
8.6	6.3	.4	8.6	7.8	1.8	5.1
4.2	6.8	10.2	2.0	5.2	3.7	5.5
5.8	9.8	2.8	8.0	8.4	4.0	
3.4	2.9	11.6	9.5	6.3	5.7	
9.3	10.9	4.3	1.3	4.4	2.4	
7.4	4.7	3.1	4.8	5.2	9.2	
1.8	3.9	5.8	9.9	7.4	5.0	

estimate limits between which the waiting times of most of the customers arriving during peak business hours would be. Also, estimate the proportion of waiting times of customers arriving during peak business hours that are less than 6 minutes.

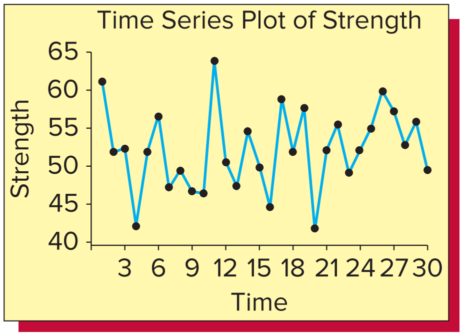
1.15 In an article entitled "Turned Off" in the June 2–4, 1995, issue of *USA Weekend*, Don Olmsted and Gigi Anders reported results of a survey where readers were invited to write in and express their opinions about sex and violence on television. The results showed that 96 percent of respondents were very or somewhat concerned about sex on TV, and 97 percent of respondents were very or somewhat concerned about violence on TV. Do you think that these results could be generalized to all television viewers in 1995? Why or why not?

1.16 THE TRASH BAG CASE[11] DS TrashBag

A company that produces and markets trash bags has developed an improved 30-gallon bag. The new bag is produced using a specially formulated plastic that is both stronger and more biodegradable than previously used plastics, and the company wishes to evaluate the strength of this bag. The *breaking strength* of a trash bag is considered to be the amount (in pounds) of a representative trash mix that when loaded into a bag suspended in the air will cause the bag to sustain significant damage (such as ripping or tearing). The company has decided to select a sample of 30 of the new trash bags. For each of 30 consecutive hours, the first trash bag produced at or after a randomly selected time during the hour is chosen. The bag is then subjected to a *breaking strength test*. The 30 breaking strengths obtained are given in Table 1.10. Using the largest and smallest observations in the data, estimate limits between which the breaking strengths of most trash bags would fall. Assume that the trash bag manufacturing process is operating consistently over time.

TABLE 1.10 Trash Bag Breaking Strengths DS TrashBag

61.1	63.8	52.1
51.8	50.5	55.4
52.2	47.3	49.1
42.1	54.5	52.1
51.9	50.0	55.0
56.5	44.6	59.9
47.1	58.9	57.2
49.4	51.9	52.7
46.7	57.7	55.8
46.5	41.7	49.5



1.5 Business Analytics and Data Mining (Optional)

LO1-10 Explain some of the uses of business analytics and data mining (Optional).

Big data, which sometimes needs quick (sometimes almost real-time) analysis for effective business decision making and which may be too massive to be analyzed by traditional statistical methods, has resulted in an extension of traditional statistics called *business analytics*. In general, **business analytics** might be defined as the use of traditional and newly developed statistical methods, advances in information systems, and techniques from *management science* to continuously and iteratively explore and investigate past business performance, with the purpose of gaining insight and improving business planning and operations. There are three broad categories of business analytics: *descriptive analytics, predictive analytics*, and *prescriptive analytics*.

Descriptive analytics

Descriptive analytics are graphical and numerical methods used to find and visualize patterns, associations, anomalies, and other relationships in data sets, with the purpose of

[11] This case is based on conversations by the authors with several employees working for a leading producer of trash bags. For purposes of confidentiality, we have withheld the company's name.

business improvement. In the next two subsections we will introduce what we call **graphical descriptive analytics** and **numerical descriptive analytics**.

Graphical descriptive analytics

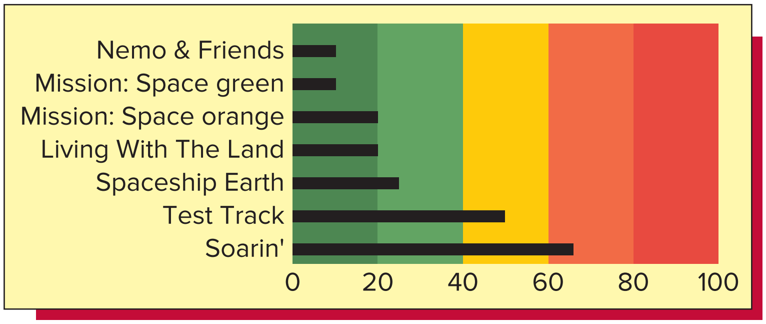
In previous examples we have illustrated using dot plots, bar charts, and time series plots to graphically display data. These and other traditional methods for displaying data are fully discussed in Sections 2.1 through 2.7 of Chapter 2. The methods discussed in these sections, and more recently developed statistical display techniques designed to take advantage of the dramatic advances in data capture, transmission, and storage, make up the toolset of *graphical descriptive analytics*. **Graphical descriptive analytics** uses the traditional and/or newer graphics to present to executives (and sometimes customers) easy-to-understand visual summaries of up-to-the minute information concerning the operational status of a business. In optional Section 2.8, we will discuss some of the new graphics, which include *gauges, bullet graphs, treemaps,* and *sparklines*. We will also see how they are used with each other and more traditional graphics to form analytic *dashboards*, which are part of *executive information systems*. As an example of one of the new graphics—the *bullet graph*—we again consider the Disney Parks Case.

C EXAMPLE 1.5 The Disney Parks Case: Predicting Ride Waiting Times

Recall that Walt Disney World Orlando collects massive amounts of data from its visitors through the MagicBands they wear. Because these data include real-time location data and the riding patterns of Disney visitors, they allow Disney to continuously predict the waiting times for each ride or attraction in its parks. This prediction is done by using the data to estimate the visitor arrival rate to the line at each ride or attraction. Then it uses statistics and a management science technique called **queuing theory** to predict the current waiting time.

One Walt Disney World Orlando park—Epcot Center—consists of the World Showcase, which features the cultures of different countries, other non-ride attractions, and (at the time that one of this book's authors visited the park) seven main rides. Near each ride, Disney posts its current predicted waiting time and also summarizes its predictions for all seven rides on large display screens located throughout the park. On February 21, 2015, one of this book's authors spent the day at Epcot and recorded the predicted waiting times (in minutes) for the seven rides. We summarize the predicted times posted at approximately 3 P.M. in the **bullet graph** in Figure 1.7. Note that the colors in the bullet graph range from dark green to red to signify short (0 to 20 minute) to very long (80 to 100 minute) predicted wait times. For each ride, a black horizontal bar representing Disney's predicted wait extends into the colors. Rather than using a bullet graph on its display screens, Disney flashes its predictions for the different rides—one at a time—on the display screens. This display method requires visitors to remember previously flashed times and thus complicates choosing the next ride. We think that Disney would be better off using a display, such as a bullet graph, that simultaneously posts all of the predicted waiting times. In whatever manner the times are displayed,

FIGURE 1.7 A Bullet Graph of Disney's Predicted Waiting Times (in minutes) for the Seven Epcot Rides Posted at 3 P.M. on February 21, 2015 DS DisneyTimes



however, they provide important information to park visitors. This is because Epcot visitors choose the three rides or attractions on which they'll gain FastPass access from certain categories, and on any given day a visitor can choose only one of Epcot's two most popular rides—Soarin' and Test Track. By viewing the posted predicted waiting times for the rides, visitors can assess whether the current predicted waiting time for a popular rides is short enough to get in line.

Finally, note that continuously monitoring predicted waiting times for the seven rides helps not only visitors but Disney management. For example, if predicted waits for rides are frequently very long, Disney might see the need to add more popular rides or attractions to its parks. As a matter of fact, Channel 13 News in Orlando reported on March 6, 2015, that Disney had announced plans to add a third "theatre" for Soarin' (a virtual ride) in order to shorten long visitor waiting times. On June 17, 2016, the third theatre for Soarin' was completed and available to park visitors.

BI

Numerical descriptive analytics

In Sections 3.1 through 3.6 of Chapter 3 we discuss methods of traditional numerical descriptive statistics—for example, using means, standard deviations, and correlation coefficients. These methods are part of the toolset of **numerical descriptive analytics,** which extends these methods and consists of the topic areas of *association learning, text mining, cluster analysis,* and *factor analysis.* We now briefly describe these topic areas, which are discussed in optional Sections 3.7 through 3.10 of Chapter 3.

Association learning This involves identifying items that tend to co-occur and finding the rules that describe their co-occurrence. For example, a supermarket chain once found that men who buy baby diapers on Thursdays also tend to buy beer on Thursdays (possibly in anticipation of watching sports on television over the weekend). This led the chain to display beer near the baby aisle in its stores. As another example, Netflix might find that customers who rent fictional dramas also tend to rent historical documentaries or that some customers will rent almost any type of movie that stars a particular actor or actress. Disney might find that visitors who spend more time at the Magic Kingdom also tend to buy Disney cartoon character clothing. Disney might also find that visitors who stay in more luxurious Disney hotels also tend to play golf on Disney courses and take cruises on the Disney Cruise Line. These types of findings are used for targeting coupons, deals, or advertising to the right potential customers.

Text mining Text mining is the science of discovering knowledge, insights, and patterns from a collection of textual documents or databases. Using *latent semantic analysis*, we can analyze the relationships between a collection of documents and the words they contain to produce a set of key concepts or factors related to the documents and words. For example, a Food and Drug Administration (FDA) administrator might use a large number of FDA drug and safety citations issued to businesses and other organizations to determine the key issues underlying the citations. Businesses and organizations can then hopefully be successfully encouraged to improve their performances in the areas related to the issues, or a company that makes pet products might analyze a set of texts from pet owners about their pets to devise advertising about their products that would appeal to the pet owners.

Cluster analysis This involves finding natural groupings, or clusters, within data without having to prespecify a set of categories. For example, Michaels Arts and Crafts might use cluster detection and customer purchase records to define different groupings of customers that reveal different buying behaviors and tastes. A financial analyst might use cluster detection to define different groupings of stocks based on the past history of stock price fluctuations. A sports marketing analyst might use cluster detection and sports fan perceptions concerning the attributes of different sports to define groupings of sports that would lead to determining why some sports are more popular than others. A waterfront developer might use cluster detection and restaurant goers' perceptions of 10 types of cultural foods to define groupings of the cultural foods that would lead to the best choice of a limited number of different cultural restaurants for the waterfront development.

Factor analysis This involves starting with a large number of correlated variables and finding fewer underlying, uncorrelated factors that describe the "essential aspects" of the large number of correlated variables. For example, a sales manager might rank job applicants on 15 attributes and then use factor analysis to find that the underlying factors that describe the essential aspects of these 15 variables are "extroverted personality," "experience," "agreeable personality," "academic ability," and "appearance." The manager of a large discount store might have shoppers rate the store on 29 attributes and use factor analysis to find that the underlying factors that describe the essential aspects of the 29 attributes are "good, friendly service," "price level," "attractiveness," spaciousness," and "size." Reducing the large number of variables to fewer underlying factors helps a business focus its activities and strategies.

It is important to note that the optional sections or descriptive analytics in Chapters 2 and 3 (Sections 2.8, 3.7, 3.8, 3.9, and 3.10) cannot only be skipped entirely but also can be read in any order without loss of continuity. Therefore, the reader has the option to choose which of these sections to study in the main flow of the course and which to study later, perhaps with Chapters 5 and 16 on predictive analytics. For readers who wish a short, easy-to-understand, and motivating introduction to descriptive analytics, we might suggest studying Sections 2.8, 3.7, and 3.8 on graphical descriptive analytics, association rules, and text mining.

Predictive analytics

Predictive analytics are methods used to predict values of a **response variable** (for example, sales of a product) on the basis of one or more **predictor variables** (for example, price and advertising expenditure). Predictive analytics determine how to make predictions by finding a relationship between previously observed values of the response variable and corresponding previously observed values of the predictor variable(s). Here, the previously observed values of the response variable are said to guide, or supervise, the learning of how to make predictions of future values of the response variable, and therefore predictive analytics are called **supervised learning** techniques. In contrast, although the descriptive analytics discussed in the previous subsection detect patterns and relationships in data, there is not a particular response variable we are trying to predict, and thus these descriptive analytics are called **unsupervised learning** techniques. The response variable we wish to predict when using predictive analytics can be quantitative (for example, sales of a product) or qualitative. If the response variable is qualitative, we use predictive analytics to perform **classification,** which assigns items to specified categories or classes. For example, a bank might study the financial records and mortgage payment histories of previous borrowers to predict whether a new mortgage applicant should be classified as a future successful or unsuccessful mortgage payer. As another example, a spam filter might use classification methodology and the differences in word usage between previously analyzed legitimate and spam e-mails to predict whether or not an incoming e-mail is spam.

Predictive analytics fall into two classes—**nonparametric predictive analytics** and **parametric predictive analytics.** Essentially, **parametric predictive analytics** find a mathematical equation that relates the response variable to the predictor variable(s) and involves unknown parameters that must be estimated and evaluated by using sample data. Parametric predictive analytics include *classical linear regression, logistic regression*, *discriminate analysis, neural networks*, and *time series forecasting*. Because evaluating the parameters in the equations obtained when using these predictive analytics requires knowledge of formal statistical inference, which is discussed in Chapters 6 through 10 (along with probability distributions) and extended in Chapters 11 through 13, we do not study parametric predictive analytics until Chapters 14 through 17. On the other hand, nonparametric predictive analytics make predictions by using a relationship between the response variable and the predictor variables that is not expressed in terms of a mathematical equation involving parameters. These analytics include *decision trees (classification and regression trees),* *k-nearest neighbors*, and *naive Bayes*' *classification*. Furthermore, the practical applications of these analytics in business can be understood with only the background of descriptive statistics and probability that is provided in Chapters 2, 3, and 4. Thus, we give an early and hopefully easy-to-understand and motivating discussion of nonparametric predictive analytics in Chapter 5. The reader thus has the option to study these analytics early in the course or nearer the time of studying parametric predictive analytics in Chapters 14 through 17.

Data mining and prescriptive analytics

Data mining is the process of discovering useful knowledge in extremely large data sets (big data). For each data mining project, data mining first uses computer science algorithms and information system techniques to extract the data needed for the project from a data source (for example, a data warehouse); data mining then uses descriptive analytics and/or predictive analytics to analyze these data. It is estimated that for any data mining project, approximately 65 percent to 90 percent of the time is spent in *data preparation*—checking, correcting, reconciling inconsistencies in, and otherwise "cleaning" the data. Also, whereas descriptive and predictive analytics might be most useful to decision makers when used with data mining, these methods can also be important, as we will see, when analyzing smaller data sets. Whatever the size of the data set being analyzed, however, it is important to turn the knowledge obtained into an optimal course of action that will lead to business improvement.

Prescriptive analytics are techniques that combine external and internal constraints (for example, the state of the economy and a company's debt relative to its equity) with results from descriptive or predictive analytics (for example, the probability that an investment being considered by the company would be successful) to recommend an optimal course of action. Prescriptive analytics include *decision theory methods* (see Chapter 19), *linear optimization, nonlinear optimization,* and *simulation.* While we will not discuss the last three of these analytics in this book (see any book on *management science* or *operations research*), we will intuitively use results from descriptive and predictive analytics to suggest business improvement courses of action.

Exercises for Section 1.5

CONCEPTS

1.17 Why are predictive analytics supervised learning techniques?

1.18 Why are descriptive analytics unsupervised learning techniques?

1.19 What is data mining?

1.20 What are prescriptive analytics?

1.6 Ratio, Interval, Ordinal, and Nominative Scales of Measurement (Optional)

LO1-11
Identify the ratio, interval, ordinal, and nominative scales of measurement (Optional).

In Section 1.1 we said that a variable is **quantitative** if its possible values are *numbers that represent quantities* (that is, "how much" or "how many"). In general, a quantitative variable is measured on a scale having a *fixed unit of measurement* between its possible values. For example, if we measure employees' salaries to the nearest dollar, then one dollar is the fixed unit of measurement between different employees' salaries. There are two types of quantitative variables: **ratio** and **interval.** A **ratio variable** is a quantitative variable measured on a scale such that ratios of its values are meaningful and there is an inherently defined zero value. Variables such as salary, height, weight, time, and distance are ratio variables. For example, a distance of zero miles is "no distance at all," and a town that is 30 miles away is "twice as far" as a town that is 15 miles away.

An **interval variable** is a quantitative variable where ratios of its values are not meaningful and there is not an inherently defined zero value. Temperature (on the Fahrenheit scale) is an interval variable. For example, zero degrees Fahrenheit does not represent "no heat at all," just that it is very cold. Thus, there is no inherently defined zero value. Furthermore, ratios of temperatures are not meaningful. For example, it makes no sense to say that 60° is twice as warm as 30°. In practice, there are very few interval variables other than temperature. Almost all quantitative variables are ratio variables.

In Section 1.1 we also said that if we simply record into which of several categories a population (or sample) unit falls, then the variable is **qualitative** (or **categorical**). There are two types of qualitative variables: **ordinal** and **nominative.** An **ordinal variable** is a qualitative variable for which there is a meaningful *ordering,* or *ranking,* of the categories. The measurements of an ordinal variable may be nonnumerical or numerical. For example, a student may be asked to rate the teaching effectiveness of a college professor as excellent, good, average, poor, or unsatisfactory. Here, one category is higher than the next one; that is, "excellent" is a higher rating than "good," "good" is a higher rating than "average," and so on. Therefore, teaching effectiveness is an ordinal variable having nonnumerical measurements. On the other hand, if (as is often done) we substitute the numbers 4, 3, 2, 1, and 0 for the ratings excellent through unsatisfactory, then teaching effectiveness is an ordinal variable having numerical measurements.

In practice, both numbers and associated words are often presented to respondents asked to rate a person or item. When numbers are used, statisticians debate whether the ordinal variable is "somewhat quantitative." For example, statisticians who claim that teaching effectiveness rated as 4, 3, 2, 1, or 0 is *not* somewhat quantitative argue that the difference between 4 (excellent) and 3 (good) may not be the same as the difference between 3 (good) and 2 (average). Other statisticians argue that as soon as respondents (students) see equally spaced numbers (even though the numbers are described by words), their responses are affected enough to make the variable (teaching effectiveness) somewhat quantitative. Generally speaking, the specific words associated with the numbers probably substantially affect whether an ordinal variable may be considered somewhat quantitative. It is important to note, however, that in practice numerical ordinal ratings are often analyzed as though they are quantitative. Specifically, various arithmetic operations (as discussed in Chapters 2 through 19) are often performed on numerical ordinal ratings. For example, a professor's teaching effectiveness average and a student's grade point average are calculated.

To conclude this section, we consider the second type of qualitative variable. A **nominative variable** is a qualitative variable for which there is no meaningful ordering, or ranking, of the categories. A person's gender, the color of a car, and an employee's state of residence are nominative variables.

Exercises for Section 1.6

CONCEPTS

1.21 Discuss the difference between a ratio variable and an interval variable.

1.22 Discuss the difference between an ordinal variable and a nominative variable.

METHODS AND APPLICATIONS

1.23 Classify each of the following qualitative variables as ordinal or nominative. Explain your answers.

Qualitative Variable	Categories
Statistics course letter grade	A B C D F
Door choice on *Let's Make A Deal*	Door #1 Door #2 Door #3
Television show classifications	TV-G TV-PG TV-14 TV-MA
Personal computer ownership	Yes No
Restaurant rating	***** **** *** ** *
Income tax filing status	Married filing jointly; Married filing separately; Single; Head of household; Qualifying widow(er)

1.24 Classify each of the following qualitative variables as ordinal or nominative. Explain your answers.

Qualitative Variable	Categories
Personal computer operating system	Windows XP; Windows Vista; Windows 7; Windows 8; Windows 10
Motion picture classifications	G PG PG-13 R NC-17 X
Level of education	Elementary; Middle school; High school; College; Graduate school
Rankings of the top 10 college football teams	1 2 3 4 5 6 7 8 9 10
Exchange on which a stock is traded	AMEX NYSE NASDAQ Other
Zip code	45056 90015 etc.

1.7 Stratified Random, Cluster, and Systematic Sampling (Optional)

LO1-12
Describe the basic ideas of stratified random, cluster, and systematic sampling (Optional).

Random sampling is not the only kind of sampling. Methods for obtaining a sample are called **sampling designs,** and the sample we take is sometimes called a **sample survey.** In this section we explain three sampling designs that are alternatives to random sampling—**stratified random sampling, cluster sampling,** and **systematic sampling.**

One common sampling design involves separately sampling important groups within a population. Then, the samples are combined to form the entire sample. This approach is the idea behind **stratified random sampling.**

> In order to select a **stratified random sample,** we divide the population into nonoverlapping groups of similar elements (people, objects, etc.). These groups are called **strata.** Then a random sample is selected from each stratum, and these samples are combined to form the full sample.

It is wise to stratify when the population consists of two or more groups that differ with respect to the variable of interest. For instance, consumers could be divided into strata based on gender, age, ethnic group, or income.

As an example, suppose that a department store chain proposes to open a new store in a location that would serve customers who live in a geographical region that consists of (1) an industrial city, (2) a suburban community, and (3) a rural area. In order to assess the potential profitability of the proposed store, the chain wishes to study the incomes of all households in the region. In addition, the chain wishes to estimate the proportion and the total number of households whose members would be likely to shop at the store. The department store chain feels that the industrial city, the suburban community, and the rural area differ with respect to income and the store's potential desirability. Therefore, it uses these subpopulations as strata and takes a stratified random sample.

Taking a stratified sample can be advantageous because such a sample takes advantage of the fact that elements in the same stratum are similar to each other. It follows that a stratified sample can provide more accurate information than a random sample of the same size. As a simple example, if all of the elements in each stratum were exactly the same, then examining only one element in each stratum would allow us to describe the entire population. Furthermore, stratification can make a sample easier (or possible) to select. Recall that, in order to take a random sample, we must have a list, or **frame** of all of the population elements. Although a frame might not exist for the overall population, a frame might exist for each stratum. For example, suppose nearly all the households in the department store's geographical region have telephones. Although there might not be a telephone directory for the overall geographical region, there might be separate telephone directories for the industrial city, the suburb, and the rural area. For more discussion of stratified random sampling, see Mendenhall, Schaeffer, and Ott (1986).

Sometimes it is advantageous to select a sample in stages. This is a common practice when selecting a sample from a very large geographical region. In such a case, a frame often does not exist. For instance, there is no single list of all registered voters in the United States. There is also no single list of all households in the United States. In this kind of situation, we can use **multistage cluster sampling.** To illustrate this procedure, suppose we wish to take a sample of registered voters from all registered voters in the United States. We might proceed as follows:

Stage 1: Randomly select a sample of counties from all of the counties in the United States.

Stage 2: Randomly select a sample of townships from each county selected in Stage 1.

Stage 3: Randomly select a sample of voting precincts from each township selected in Stage 2.

Stage 4: Randomly select a sample of registered voters from each voting precinct selected in Stage 3.

We use the term *cluster sampling* to describe this type of sampling because at each stage we "cluster" the voters into subpopulations. For instance, in Stage 1 we cluster the voters into counties, and in Stage 2 we cluster the voters in each selected county into townships. Also, notice that the random sampling at each stage can be carried out because there are lists of (1) all counties in the United States, (2) all townships in each county, (3) all voting precincts in each township, and (4) all registered voters in each voting precinct.

As another example, consider sampling the households in the United States. We might use Stages 1 and 2 above to select counties and townships within the selected counties. Then, if there is a telephone directory of the households in each township, we can randomly sample households from each selected township by using its telephone directory. Because *most* households today have telephones, and telephone directories are readily available, most national polls are now conducted by telephone. Further, polling organizations have recognized that many households are giving up landline phones, and have developed ways to sample households that only have cell phones.

It is sometimes a good idea to combine stratification with multistage cluster sampling. For example, suppose a national polling organization wants to estimate the proportion of all registered voters who favor a particular presidential candidate. Because the presidential preferences of voters might tend to vary by geographical region, the polling organization might divide the United States into regions (say, Eastern, Midwestern, Southern, and Western regions). The polling organization might then use these regions as strata, and might take a multistage cluster sample from each stratum (region).

The analysis of data produced by multistage cluster sampling can be quite complicated. For a more detailed discussion of cluster sampling, see Mendenhall, Schaeffer, and Ott (1986).

In order to select a random sample, we must number the elements in a frame of all the population elements. Then we use a random number table (or a random number generator on a computer) to make the selections. However, numbering all the population elements can be quite time-consuming. Moreover, random sampling is used in the various stages of many complex sampling designs (requiring the numbering of numerous populations). Therefore, it is useful to have an alternative to random sampling. One such alternative is called **systematic sampling.** In order to systematically select a sample of n elements without replacement from a frame of N elements, we divide N by n and round the result down to the nearest whole number. Calling the rounded result ℓ, we then randomly select one element from the first ℓ elements in the frame—this is the first element in the systematic sample. The remaining elements in the sample are obtained by selecting every ℓth element following the first (randomly selected) element. For example, suppose we wish to sample a population of $N = 14{,}327$ allergists to investigate how often they have prescribed a particular drug during the last year. A medical society has a directory listing the 14,327 allergists, and we wish to draw a systematic sample of 500 allergists from this frame. Here we compute 14,327/500 = 28.654, which is 28 when rounded down. Therefore, we number the first 28 allergists in the directory from 1 to 28, and we use a random number table to randomly select one of the first 28 allergists. Suppose we select allergist number 19. We interview allergist 19 and every 28th allergist in the frame thereafter, so we choose allergists 19, 47, 75, and so forth until we obtain our sample of 500 allergists. In this scheme, we must number the first 28 allergists, but we do not have to number the rest because we can "count off" every 28th allergist in the directory. Alternatively, we can measure the approximate amount of space in the directory that it takes to list 28 allergists. This measurement can then be used to select every 28th allergist.

Exercises for Section 1.7

CONCEPTS

1.25 When is it appropriate to use stratified random sampling? What are strata, and how should strata be selected?

1.26 When is cluster sampling used? Why do we describe this type of sampling by using the term *cluster*?

1.27 Explain how to take a systematic sample of 100 companies from the 1,853 companies that are members of an industry trade association.

1.28 Explain how a stratified random sample is selected. Discuss how you might define the strata to survey student opinion on a proposal to charge all students a $100 fee for a new university-run bus system that will provide transportation between off-campus apartments and campus locations.

1.29 Marketing researchers often use city blocks as clusters in cluster sampling. Using this fact, explain how a market researcher might use multistage cluster sampling to select a sample of consumers from all cities having a population of more than 10,000 in a large state having many such cities.

1.8 More about Surveys and Errors in Survey Sampling (Optional)

LO1-13
Describe basic types of survey questions, survey procedures, and sources of error (Optional).

We have seen in Section 1.2 that people in surveys are asked questions about their behaviors, opinions, beliefs, and other characteristics. In this section we discuss various issues related to designing surveys and the errors that can occur in survey sampling.

Types of survey questions

Survey instruments can use **dichotomous** ("yes or no"), **multiple-choice,** or **open-ended** questions. Each type of question has its benefits and drawbacks. Dichotomous questions are usually clearly stated, can be answered quickly, and yield data that are easily analyzed. However, the information gathered may be limited by this two-option format. If we limit voters to expressing support or disapproval for stem-cell research, we may not learn the nuanced reasoning that voters use in weighing the merits and moral issues involved. Similarly, in today's heterogeneous world, it would be unusual to use a dichotomous question to categorize a person's religious preferences. Asking whether respondents are Christian or non-Christian (or to use any other two categories like Jewish or non-Jewish; Muslim or non-Muslim) is certain to make some people feel their religion is being slighted. In addition, this is a crude and unenlightening way to learn about religious preferences.

Multiple-choice questions can assume several different forms. Sometimes respondents are asked to choose a response from a list (for example, possible answers to the religion question could be Jewish, Christian, Muslim, Hindu, Agnostic, or Other). Other times, respondents are asked to choose an answer from a numerical range. We could ask the question:

"In your opinion, how important are SAT scores to a college student's success?"

Not important at all 1 2 3 4 5 Extremely important

These numerical responses are usually summarized and reported in terms of the average response, whose size tells us something about the perceived importance. The Zagat restaurant survey (www.zagat.com) asks diners to rate restaurants' food, décor, and service, each on a scale of 1 to 30 points, with a 30 representing an incredible level of satisfaction. Although the Zagat scale has an unusually wide range of possible ratings, the concept is the same as in the more common 5-point scale.

Open-ended questions typically provide the most honest and complete information because there are no suggested answers to divert or bias a person's response. This kind of question is often found on instructor evaluation forms distributed at the end of a college course. College students at Georgetown University are asked the open-ended question, "What comments would you give to the instructor?" The responses provide the instructor feedback that may be missing from the initial part of the teaching evaluation survey, which consists of numerical multiple-choice ratings of various aspects of the course. While these numerical ratings can be used to compare instructors and courses, there are no easy comparisons of the diverse responses instructors receive to the open-ended question. In fact, these responses are often seen only by the instructor and are useful, constructive tools for the teacher despite the fact they cannot be readily summarized.

Survey questionnaires must be carefully constructed so they do not inadvertently bias the results. Because survey design is such a difficult and sensitive process, it is not uncommon for a pilot survey to be taken before a lot of time, effort, and financing go into collecting a large amount of data. Pilot surveys are similar to the beta version of a new electronic product; they are tested out with a smaller group of people to work out the "kinks" before being used on a larger scale. Determination of the sample size for the final survey is an important process for many reasons. If the sample size is too large, resources may be wasted during the data collection. On the other hand, not collecting enough data for a meaningful analysis will obviously be detrimental to the study. Fortunately, there are several formulas that will help decide how large a sample should be, depending on the goal of the study and various other factors.

Types of surveys

There are several different survey types, and we will explore just a few of them. The **phone survey** is particularly well-known (and often despised). A phone survey is inexpensive and usually conducted by callers who have very little training. Because of this and the impersonal nature of the medium, the respondent may misunderstand some of the questions. A further drawback is that some people cannot be reached and that others may refuse to answer some or all of the questions. Phone surveys are thus particularly prone to have a low **response rate.**

> The **response rate** is the proportion of all people whom we attempt to contact that actually respond to a survey. A low response rate can destroy the validity of a survey's results.

It can be difficult to collect good data from unsolicited phone calls because many of us resent the interruption. The calls often come at inopportune times, intruding on a meal or arriving just when we have climbed a ladder with a full can of paint. No wonder we may fantasize about turning the tables on the callers and calling *them* when it is least convenient.

Numerous complaints have been filed with the Federal Trade Commission (FTC) about the glut of marketing and survey telephone calls to private residences. The National Do Not Call Registry was created as the culmination of a comprehensive, three-year review of the Telemarketing Sales Rule (TSR) (www.ftc.gov/donotcall/). This legislation allows people to enroll their phone numbers on a website so as to prevent most marketers from calling them.

Self-administered surveys, or **mail surveys,** are also very inexpensive to conduct. However, these also have their drawbacks. Often, recipients will choose not to reply unless they receive some kind of financial incentive or other reward. Generally, after an initial mailing, the response rate will fall between 20 and 30 percent. Response rates can be raised with successive follow-up reminders, and after three contacts, they might reach between 65 and 75 percent. Unfortunately, the entire process can take significantly longer than a phone survey would.

Web-based surveys have become increasingly popular, but they suffer from the same problems as mail surveys. In addition, as with phone surveys, respondents may record their true reactions incorrectly because they have misunderstood some of the questions posed.

A personal interview provides more control over the survey process. People selected for interviews are more likely to respond because the questions are being asked by someone face-to-face. Questions are less likely to be misunderstood because the people conducting the interviews are typically trained employees who can clear up any confusion arising during the process. On the other hand, interviewers can potentially "lead" a respondent by body language, which signals approval or disapproval of certain sorts of answers. They can also prompt certain replies by providing too much information. **Mall surveys** are examples of personal interviews. Interviewers approach shoppers as they pass by and ask them to answer the survey questions. Response rates around 50 percent are typical. Personal interviews are more costly than mail or phone surveys. Obviously, the objective of the study will be important in deciding upon the survey type employed.

Errors occurring in surveys

In general, the goal of a survey is to obtain accurate information from a group, or sample, that is representative of the entire population of interest. We are trying to estimate some aspect (numerical descriptor) of the entire population from a subset of the population. This is not an easy task, and there are many pitfalls. First and foremost, the *target population* must be well defined and a *sample frame* must be chosen.

> The **target population** is the entire population of interest to us in a particular study.

Are we intending to estimate the average starting salary of students graduating from any college? Or from four-year colleges? Or from business schools? Or from a particular business school?

The **sample frame** is a list of sampling elements (people or things) from which the sample will be selected. It should closely agree with the target population.

Consider a study to estimate the average starting salary of students who have graduated from the business school at Miami University of Ohio over the last five years; the target population is obviously that particular group of graduates. A sample frame could be the Miami University Alumni Association's roster of business school graduates for the past five years. Although it will not be a perfect replication of the target population, it is a reasonable frame.

We now discuss two general classes of survey errors: **errors of nonobservation** and **errors of observation.** From the sample frame, units are randomly chosen to be part of the sample. Simply by virtue of the fact that we are taking a sample instead of a census, we are susceptible to *sampling error.*

Sampling error is the difference between a numerical descriptor of the population and the corresponding descriptor of the sample.

Sampling error occurs because our information is incomplete. We observe only the portion of the population included in the sample while the remainder is obscured. Suppose, for example, we wanted to know about the heights of 13-year-old boys. There is extreme variation in boys' heights at this age. Even if we could overcome the logistical problems of choosing a random sample of 20 boys, there is nothing to guarantee the sample will accurately reflect heights at this age. By sheer luck of the draw, our sample could include a higher proportion of tall boys than appears in the population. We would then overestimate average height at this age (to the chagrin of the shorter boys). Although samples tend to look more similar to their parent populations as the sample sizes increase, we should always keep in mind that sample characteristics and population characteristics are not the same.

If a sample frame is not identical to the target population, we will suffer from an *error of coverage*.

Undercoverage occurs when some population elements are excluded from the process of selecting the sample.

Undercoverage was part of the problem dooming the *Literary Digest* Poll of 1936. Although millions of Americans were included in the poll, the large sample size could not rescue the poll results. The sample represented those who could afford phone service and magazine subscriptions in the lean Depression years, but in excluding everyone else, it failed to yield an honest picture of the entire American populace. Undercoverage often occurs when we do not have a complete, accurate list of all the population elements. If we select our sample from an incomplete list, like a telephone directory or a list of all Internet subscribers in a region, we automatically eliminate those who cannot afford phone or Internet service. Even today, 7 to 8 percent of the people in the United States do not own telephones. Low-income people are often underrepresented in surveys. If underrepresented groups differ from the rest of the population with respect to the characteristic under study, the survey results will be biased.

Often, pollsters cannot find all the people they intend to survey, and sometimes people who are found will refuse to answer the questions posed. Both of these are examples of the **nonresponse** problem. Unfortunately, there may be an association between how difficult it is to find and elicit responses from people and the type of answers they give.

Nonresponse occurs whenever some of the individuals who were supposed to be included in the sample are not.

For example, universities often conduct surveys to learn how graduates have fared in the workplace. The alumnus who has risen through the corporate ranks is more likely to have a current address on file with his alumni office and to be willing to share career information than a classmate who has foundered professionally. We should be politely skeptical about reports touting the average salaries of graduates of various university programs. In some surveys, 35 percent or more of the selected individuals cannot be contacted—even when several callbacks are made. In such cases, other participants are often substituted for those who cannot be contacted. If the substitutes and the originally selected participants differ with respect to the characteristic under study, the survey will be biased. Furthermore, people who will answer highly sensitive, personal, or embarrassing questions might be very different from those who will not.

As discussed in Section 1.4, the opinions of those who bother to complete a voluntary response survey may be dramatically different from those who do not. (Recall the Ann Landers question about having children.) The viewer voting on the television show *American Idol* is another illustration of **selection bias,** because only those who are interested in the outcome of the show will bother to phone in or text message their votes. The results of the voting are not representative of the performance ratings the country would give as a whole.

Errors of observation occur when data values are recorded incorrectly. Such errors can be caused by the data collector (the interviewer), the survey instrument, the respondent, or the data collection process. For instance, the manner in which a question is asked can influence the response. Or, the order in which questions appear on a questionnaire can influence the survey results. Or, the data collection method (telephone interview, questionnaire, personal interview, or direct observation) can influence the results. A **recording error** occurs when either the respondent or interviewer incorrectly marks an answer. Once data are collected from a survey, the results are often entered into a computer for statistical analysis. When transferring data from a survey form to a spreadsheet program like Excel, Minitab, or JMP, there is potential for entering them incorrectly. Before the survey is administered, the questions need to be very carefully worded so that there is little chance of misinterpretation. A poorly framed question might yield results that lead to unwarranted decisions. Scaled questions are particularly susceptible to this type of error. Consider the question "How would you rate this course?" Without a proper explanation, the respondent may not know whether "1" or "5" is the best.

If the survey instrument contains highly sensitive questions and respondents feel compelled to answer, they may not tell the truth. This is especially true in personal interviews. We then have what is called **response bias.** A surprising number of people are reluctant to be candid about what they like to read or watch on television. People tend to overreport "good" activities like reading respected newspapers and underreport their "bad" activities like delighting in the *National Enquirer*'s stories of alien abductions and celebrity meltdowns. Imagine, then, the difficulty in getting honest answers about people's gambling habits, drug use, or sexual histories. Response bias can also occur when respondents are asked slanted questions whose wording influences the answer received. For example, consider the following question:

Which of the following best describes your views on gun control?

1. The government should take away our guns, leaving us defenseless against heavily armed criminals.
2. We have the right to keep and bear arms.

This question is biased toward eliciting a response against gun control.

Exercises for Section 1.8

CONCEPTS

1.30 Explain:

a Three types of surveys and discuss their advantages and disadvantages.

b Three types of survey questions and discuss their advantages and disadvantages.

1.31 Explain each of the following terms:

a Undercoverage. **b** Nonresponse. **c** Response bias.

1.32 A market research firm sends out a web-based survey to assess the impact of advertisements placed on a search engine's results page. About 65 percent of the surveys were answered and sent back. What types of errors are possible in this scenario?

Chapter Summary

We began this chapter by discussing **data.** We learned that the data that are collected for a particular study are referred to as a **data set,** and we learned that **elements** are the entities described by a data set. In order to determine what information we need about a group of elements, we define important **variables,** or characteristics, describing the elements. **Quantitative variables** are variables that use numbers to measure quantities (that is, "how much" or "how many") and **qualitative, or categorical, variables** simply record into which of several categories an element falls.

We next discussed the difference between cross-sectional data and time series data. **Cross-sectional data** are data collected at the same or approximately the same point in time. **Time series data** are data collected over different time periods, and we saw that time series data are often depicted by using a **time series plot.**

Next we learned about data sources. **Primary data** are collected by an individual through personally planned experimentation or observation, while **secondary data** are taken from existing sources. We discussed some readily available existing data sources, and we learned the difference between *experimental* and *observational* studies. We found that a study is **experimental** when we are able to set or manipulate the factors that may be related to the response variable and that a study is **observational** when we are unable to control the factors of interest. We learned that with the increased use of online purchasing and with increased competition, businesses have become more aggressive about collecting information about customer transactions. Dramatic advances in data capture, data transmission, and data storage capabilities are enabling organizations to integrate various databases into **data warehouses.** The term **big data** refers to massive amounts of data, often collected at very fast rates in real time and in different forms and sometimes needing quick preliminary analysis for effective business decision making.

We often collect data to study a **population,** which is the set of all elements about which we wish to draw conclusions. We saw that, because many populations are too large to examine in their entirety, we frequently study a population by selecting a **sample,** which is a subset of the population elements. We saw that we often wish to describe a population or sample and that **descriptive statistics** is the science of describing the important aspects of a population or sample. We also learned that if a population is large and we need to select a sample from it, we use what is called **statistical inference,** which is the science of using a sample to make generalizations about the important aspects of a population.

Next we learned that if the information contained in a sample is to accurately represent the population, then the sample should be **randomly selected** from the population. In Section 1.4 we formally defined a **random sample,** and we studied three cases that introduced how we can take a random (or approximately random) sample. The methods we illustrated included using a **random number table** and using **computer-generated random numbers.** We also learned that we often wish to sample a **process** over time, and we illustrated how such sampling might be done. Finally, Section 1.4 presented some ethical guidelines for statistical practice.

We concluded this chapter with four optional sections. In optional Section 1.5 we learned that big data has resulted in an extension of traditional statistics called **business analytics**, and we introduced some basic ideas of **descriptive analytics, predictive analytics, data mining,** and **prescriptive analytics**. In optional Section 1.6, we considered different types of quantitative and qualitative variables. We learned that there are two types of **quantitative variables—ratio variables,** which are measured on a scale such that ratios of its values are meaningful and there is an inherently defined zero value, and **interval variables,** for which ratios are not meaningful and there is no inherently defined zero value. We also saw that there are two types of **qualitative variables—ordinal variables,** for which there is a meaningful ordering of the categories, and **nominative variables,** for which there is no meaningful ordering of the categories. Optional Section 1.7 introduced several advanced sampling designs: **stratified random sampling, cluster sampling,** and **systematic sampling.** Finally, optional Section 1.8 discussed more about sample surveys. Topics included **types of surveys** (such as phone surveys, mail surveys, and mall surveys), types of **survey questions** (such as dichotomous, multiple-choice, and open-ended questions), and **survey errors** (such as sampling error, error due to undercoverage, and error due to nonresponse).

Glossary of Terms

big data: Massive amounts of data, often collected at very fast rates in real time and in different forms and sometimes needing quick preliminary analysis for effective business decision making.

business analytics: The use of traditional and newly developed statistical methods, advances in information systems, and techniques from *management science* to continuously and iteratively explore and investigate past business performance, with the purpose of gaining insight and improving business planning and operations.

categorical (qualitative) variable: A variable having values that indicate into which of several categories a population element belongs.

census: An examination of all the elements in a population.

cluster sampling (multistage cluster sampling): A sampling design in which we sequentially cluster population elements into subpopulations.