14TH EDITION

STATISTICS

FOR BUSINESS AND ECONOMICS

McClave
Benson
Sincich

This page intentionally left blank

APPLET CORRELATION

Applet	Concept Illustrated	Description	Applet Activity
Random numbers	Uses a random number generator to determine the experimental units to be included in a sample.	Generates random numbers from a range of integers specified by the user.	**1.1**, 25; **1.2**, 25; **3.6**, 180; **4.1**, 200; **4.2**, 200; **4.8**, 247
Sample from a population	Assesses how well a sample represents the population and the role that sample size plays in the process.	Produces random sample from population from specified sample size and population distribution shape. Reports mean, median, and standard deviation; applet creates plot of sample.	**4.4**, 214; **4.6**, 238; **4.7**, 252
Sampling distributions	Compares means and standard deviations of distributions; assesses effect of sample size; illustrates undbiasedness.	Simulates repeatedly choosing samples of a fixed size n from a population with specified sample size, number of samples, and shape of population distribution. Applet reports means, medians, and standard deviations; creates plots for both.	**5.1**, 290; **5.2**, 290

Long-run probability demonstrations illustrate the concept that theoretical probabilities are long-run experimental probabilities.

Applet	Concept Illustrated	Description	Applet Activity
Simulating probability of rolling a 6	Investigates relationship between theoretical and experimental probabilities of rolling 6 as number of die rolls increases.	Reports and creates frequency histogram for each outcome of each simulated roll of a fair die. Students specify number of rolls; applet calculates and plots proportion of 6s.	**3.1**, 144; **3.3**, 156; **3.4**, 157; **3.5**, 170
Simulating probability of rolling a 3 or 4	Investigates relationship between theoretical and experimental probabilities of rolling 3 or 4 as number of die rolls increases.	Reports outcome of each simulated roll of a fair die; creates frequency histogram for outcomes. Students specify number of rolls; applet calculates and plots proportion of 3s and 4s.	**3.3**, 156; **3.4**, 157
Simulating the probability of heads: fair coin	Investigates relationship between theoretical and experimental probabilities of getting heads as number of fair coin flips increases.	Reports outcome of each fair coin flip and creates a bar graph for outcomes. Students specify number of flips; applet calculates and plots proportion of heads.	**3.2**, 144; **4.2**, 200
Simulating probability of heads: unfair coin $(P(H) = .2)$	Investigates relationship between theoretical and experimental probabilities of getting heads as number of unfair coin flips increases.	Reports outcome of each flip for a coin where heads is less likely to occur than tails and creates a bar graph for outcomes. Students specify number of flips; applet calculates and plots the proportion of heads.	**4.3**, 214
Simulating probability of heads: unfair coin $(P(H) = .8)$	Investigates relationship between theoretical and experimental probabilities of getting heads as number of unfair coin flips increases.	Reports outcome of each flip for a coin where heads is more likely to occur than tails and creates a bar graph for outcomes. Students specify number of flips; applet calculates and plots the proportion of heads.	**4.3**, 214
Simulating the stock market	Theoretical probabilities are long run experimental probabilities.	Simulates stock market fluctuation. Students specify number of days; applet reports whether stock market goes up or down daily and creates a bar graph for outcomes. Calculates and plots proportion of simulated days stock market goes up.	**4.5**, 215
Mean versus median	Investigates how skewedness and outliers affect measures of central tendency.	Students visualize relationship between mean and median by adding and deleting data points; applet automatically updates mean and median.	**2.1**, 70; **2.2**, 70; **2.3**, 70

(*Continued*)

Applet	Concept Illustrated	Description	Applet Activity
Standard deviation	Investigates how distribution shape and spread affect standard deviation.	Students visualize relationship between mean and standard deviation by adding and deleting data points; applet updates mean and standard deviation.	**2.4**, 79; **2.5**, 79; **2.6**, 79; **2.7**, 101
Confidence intervals for a mean (the impact of confidence level)	Not all confidence intervals contain the population mean. Investigates the meaning of 95% and 99% confidence.	Simulates selecting 100 random samples from population; finds 95% and 99% confidence intervals for each. Students specify sample size, distribution shape, and population mean and standard deviation; applet plots confidence intervals and reports number and proportion containing true mean.	**6.1**, 314; **6.2**, 314
Confidence intervals for a mean (not knowing standard deviation)	Confidence intervals obtained using the sample standard deviation are different from those obtained using the population standard deviation. Investigates effect of not knowing the population standard deviation.	Simulates selecting 100 random samples from the population and finds the 95% z-interval and 95% t-interval for each. Students specify sample size, distribution shape, and population mean and standard deviation; applet plots confidence intervals and reports number and proportion containing true mean.	**6.3**, 324; **6.4**, 324
Confidence intervals for a proportion	Not all confidence intervals contain the population proportion. Investigates the meaning of 95% and 99% confidence.	Simulates selecting 100 random samples from the population and finds the 95% and 99% confidence intervals for each. Students specify population proportion and sample size; applet plots confidence intervals and reports number and proportion containing true proportion.	**6.5**, 332; **6.6**, 332
Hypothesis tests for a mean	Not all tests of hypotheses lead correctly to either rejecting or failing to reject the null hypothesis. Investigates the relationship between the level of confidence and the probabilities of making Type I and Type II errors.	Simulates selecting 100 random samples from population; calculates and plots t statistic and P-value for each. Students specify population distribution shape, mean, and standard deviation; sample size, and null and alternative hypotheses; applet reports number and proportion of times null hypothesis is rejected at both 0.05 and 0.01 levels.	**7.1**, 373; **7.2**, 384; **7.3**, 384; **7.4**, 384
Hypothesis tests for a proportion	Not all tests of hypotheses lead correctly to either rejecting or failing to reject the null hypothesis. Investigates the relationship between the level of confidence and the probabilities of making Type I and Type II errors.	Simulates selecting 100 random samples from population; calculates and plots z-statistic and P-value for each. Students specify population proportion, sample size, and null and alternative hypotheses; applet reports number and proportion of times null hypothesis is rejected at 0.05 and 0.01 levels.	**7.5**, 400; **7.6**, 401
Correlation by eye	Correlation coefficient measures strength of linear relationship between two variables. Teaches user how to assess strength of a linear relationship from a scattergram.	Computes correlation coefficient r for a set of bivariate data plotted on a scattergram. Students add or delete points and guess value of r; applet compares guess to calculated value.	**11.2**, 654
Regression by eye	The least squares regression line has a smaller SSE than any other line that might approximate a set of bivariate data. Teaches students how to approximate the location of a regression line on a scattergram.	Computes least squares regression line for a set of bivariate data plotted on a scattergram. Students add or delete points and guess location of regression line by manipulating a line provided on the scattergram; applet plots least squares line and displays the equations and the SSEs for both lines.	**11.1**, 629

Statistics

for Business and Economics

This page intentionally left blank

EDITION 14

Statistics

for Business and Economics

James T. **MCCLAVE**
Info Tech, Inc.
University of Florida

P. George **BENSON**
College of Charleston

Terry **SINCICH**
University of South Florida

Pearson

Content Development: Robert Carroll
Content Management: Suzanna Bainbridge, Amanda Brands
Content Production: Jean Choe, Peggy McMahon
Product Management: Karen Montgomery
Product Marketing: Alicia Wilson
Rights and Permissions: Tanvi Bhatia/Anjali Singh
Please contact https://support.pearson.com/getsupport/s/ with any queries on this content

Acknowledgments of third-party content appear on page C-1, which constitutes an extension of this copyright page.

Library of Congress Control Number: 2020919332

ScoutAutomatedPrintCode

Rental
ISBN-10: 0-13-685535-0
ISBN-13: 978-0-13-685535-4

Contents

Preface

This 14th edition of *Statistics for Business and Economics* is an introductory text emphasizing inference, with extensive coverage of data collection and analysis as needed to evaluate the reported results of statistical studies and make good decisions. As in earlier editions, the text stresses the development of statistical thinking, the assessment of credibility and value of the inferences made from data, both by those who consume and by those who produce them. It assumes a mathematical background of basic algebra.

The text incorporates the following features, developed from the American Statistical Association (ASA) sponsored conferences on *Making Statistics More Effective in Schools of Business* (MSMESB) and ASA's Guidelines for Assessment and Instruction in Statistics Education (GAISE) Project:

- Emphasize statistical literacy and develop statistical thinking
- Use real data in applications
- Use technology for developing conceptual understanding and analyzing data
- Foster active learning in the classroom
- Stress conceptual understanding rather than mere knowledge of procedures
- Emphasize intuitive concepts of probability

New in the 14th Edition

- **More than 1,200 exercises, with revisions and updates to 30%.** Many new and updated exercises, based on contemporary business-related studies and real data, have been added. Most of these exercises foster critical thinking skills.
- **Data Informed Development.** The authors analyzed aggregated student usage and performance data from MyLab Statistics for the previous edition of this text. The results of this analysis helped improve the quality and quantity of exercises that matter most to instructors and students.
- **Updated technology.** All printouts from statistical software (Excel 2019/XLSTAT, StatCrunch 3.0, Minitab 19, and the TI-84 Graphing Calculator) and corresponding instructions for use have been revised to reflect the latest versions of the software.
- **Statistics in Action Cases Updated.** Three of the 14 Statistics in Action cases have been updated. All cases are based on real data from a recent business study.
- **Continued Emphasis on Ethics.** Where appropriate, boxes have been added to emphasize the importance of ethical behavior when collecting, analyzing, and interpreting data with statistics.
- **Business Analytics.** The importance of statistical thinking to successful business analytics is established early in the text.
- **Short Video Tutorials.** New videos guide students through real-life applications of chapter topics to illustrate how these concepts translate to everyday life.

Hallmark Strengths

We have maintained the pedagogical features of *Statistics for Business and Economics* that we believe make it unique among introductory business statistics texts. These features, which assist the student in achieving an overview of statistics and an understanding of its relevance in both the business world and everyday life, are as follows:

- **Use of Examples as a Teaching Device** Almost all new ideas are introduced and illustrated by data-based applications and examples. We believe that students better understand definitions, generalizations, and theoretical concepts *after* seeing an application. All examples have three components: (1) "Problem," (2) "Solution," and (3) "Look Back" (or "Look Ahead"). This step-by-step process provides students with a defined structure by which to approach problems and enhances their problem-solving skills. The "Look Back/Look Ahead" feature often gives helpful hints to solving the problem and/or provides a further reflection or insight into the concept or procedure that is covered.
- **Now Work** A "Now Work" exercise suggestion follows each example. The Now Work exercise (marked with the NW icon in the exercise sets) is similar in style and concept to the text example. This provides students with an opportunity to immediately test and confirm their understanding.
- **Statistics in Action** Each chapter begins with a case study based on an actual contemporary, controversial or high-profile issue in business. Relevant research questions and data from the study are presented and the proper analysis is demonstrated in short "Statistics in Action Revisited" sections throughout the chapter. These motivate students to critically evaluate the findings and think through the statistical issues involved.
- **"Hands-On" Activities for Students** At the end of each chapter, students are provided with an opportunity to participate in hands-on classroom activities, ranging from real data collection to formal statistical analysis. These activities are designed to be performed by students individually or as a class.
- **Applet Exercises.** The text is accompanied by applets (short computer programs), available on the student resource site and in MyLab Statistics. These point-and-click applets allow students to easily run simulations that visually demonstrate some of the more difficult statistical concepts (e.g., sampling distributions and confidence intervals.) Each chapter contains several optional applet exercises in the exercise sets. They are denoted with the following Applet icon: .
- **Real-World Business Cases** Seven extensive business problem-solving cases, with real data and assignments for the student, are provided. Each case serves as a good capstone and review of the material that has preceded it. Typically, these cases follow a group of two or three chapters and require the student to apply the methods presented in these chapters.
- **Real Data–Based Exercises** The text includes more than 1,200 exercises based on applications in a variety of business disciplines and research areas. All applied exercises use current real data extracted from current publications (e.g., newspapers, magazines, current journals, and the Internet). Some students have difficulty learning the mechanics of statistical techniques when all problems are couched in terms of realistic applications. For this reason, all exercise sections are divided into at least four parts:

  Learning the Mechanics. Designed as straightforward applications of new concepts, these exercises allow students to test their ability to comprehend a mathematical concept or a definition.

  Applying the Concepts—Basic. Based on applications taken from a wide variety of business journals, newspapers, and other sources, these short exercises help students to begin developing the skills necessary to diagnose and analyze real-world problems.

  Applying the Concepts—Intermediate. Based on more detailed real-world applications, these exercises require students to apply their knowledge of the technique presented in the section.

Applying the Concepts—Advanced. These more difficult real-data exercises require students to use their critical thinking skills.

Critical Thinking Challenges. Placed at the end of the "Supplementary Exercises" section only, this feature presents students with one or two challenging business problems.

- **Exploring Data with Statistical Computer Software and the Graphing Calculator** Each statistical analysis method presented is demonstrated using output from three leading Windows-based statistical software packages: Excel/XLSTAT, StatCrunch, and Minitab. Students are exposed early and often to computer printouts they will encounter in today's hi-tech business world.
- **"Using Technology" Tutorials** At the end of each chapter are statistical software tutorials with point-and-click instructions (with screen shots) for Minitab, StatCrunch, and Excel/XLSTAT. These tutorials are easily located and show students how to best use and maximize statistical software. In addition, output and keystroke instructions for the TI-84 Graphing Calculator are presented.
- **Profiles of Statisticians in History (Biography)** Brief descriptions of famous statisticians and their achievements are presented in side boxes. In reading these profiles, students will develop an appreciation for the statistician's efforts and the discipline of statistics as a whole.
- **Data and Applets** The text is accompanied by a website (www.pearsonhighered.com/mathstatsresources/) that contains files for all of the data sets marked with an icon in the text. These include data sets for text examples, exercises, Statistics in Action, and Real-World cases. All data files are saved in multiple formats: Excel and Minitab. This website also contains the applets that are used to illustrate statistical concepts.

Flexibility in Coverage

The text is written to allow the instructor flexibility in coverage of topics. Suggestions for two topics, probability and regression, are given below.

- **Probability and Counting Rules** One of the most troublesome aspects of an introductory statistics course is the study of probability. Probability poses a challenge for instructors because they must decide on the level of presentation, and students find it a difficult subject to comprehend. We believe that one cause for these problems is the mixture of probability and counting rules that occurs in most introductory texts. Consequently, we have included the counting rules (with examples) in an appendix (Appendix B) rather than in the body of Chapter 3. Thus, the instructor can control the level of coverage of probability.
- **Multiple Regression and Model Building** This topic represents one of the most useful statistical tools for the solution of applied problems. Although an entire text could be devoted to regression modeling, we feel that we have presented coverage that is understandable, usable, and much more comprehensive than the presentations in other introductory statistics texts. We devote two full chapters to discussing the major types of inferences that can be derived from a regression analysis, showing how these results appear in the output from statistical software, and, most important, selecting multiple regression models to be used in an analysis. Thus, the instructor has the choice of a one-chapter coverage of simple linear regression (Chapter 11), a two-chapter treatment of simple and multiple regression (excluding the sections on model building in Chapter 12), or complete coverage of regression analysis, including model building and regression diagnostics. This extensive coverage of such useful statistical tools will provide added evidence to the student of the relevance of statistics to real-world problems.
- **Role of Calculus in Footnotes** Although the text is designed for students with a non-calculus background, **footnotes** explain the role of calculus in various derivations. Footnotes are also used to inform the student about some of the theory underlying certain methods of analysis. These footnotes allow additional flexibility in the mathematical and theoretical level at which the material is presented.

Acknowledgments

This book reflects the efforts of a great many people over a number of years. First, we would like to thank the following professors, whose reviews and comments on this and prior editions have contributed to the 14th edition:

Reviewers of the 14th Edition of *Statistics for Business and Economics*

Anna Errore, University of Minnesota Twin Cities
Alka Gandhi, University of Maryland-College Park
Stacey Hachigian, Humber College
Seunghee Lee, Pellissippi State Community College
Amit Mitra, Auburn University
Yu Yue, Baruch College

Reviewers of Previous Editions

ALABAMA Volodymyr Melnykov, *University of Alabama–Tuscaloosa* **ARKANSAS** Julie Trivitt, *University of Arkansas* **CALIFORNIA** Joyce Curley-Daly, Jim Daly, Robert K. Smidt, *California Polytechnic State University* • Jim Davis, *Golden Gate University* • Carol Eger, *Stanford University* • Paul W. Guy, *California State University, Chico* • Judd Hammack, P. Kasliwal, *California State University, Los Angeles* • Mabel T. King, *California State University, Fullerton* • James Lackritz, *California State University, San Diego* • Beth Rose, *University of Southern California* • Daniel Sirvent, *Vanguard University* **COLORADO** Rick L. Edgeman, Charles F. Warnock, *Colorado State University* • Eric Huggins, *Fort Lewis College* • William J. Weida, *United States Air Force Academy* **CONNECTICUT** Alan E. Gelfand, Joseph Glaz, Timothy J. Killeen, *University of Connecticut* **DELAWARE** Christine Ebert, *University of Delaware* **DISTRICT OF COLUMBIA** Phil Cross, Jose Luis Guerrero-Cusumano, *Georgetown University* • Gaminie Meepagala, *Howard University* **FLORIDA** John M. Charnes, *University of Miami* • C. Brad Davis, *Clearwater Christian College* • Vivian Jones, *Bethune-Cookman University* • Fred Leysieffer, Pi-Erh Lin, Doug Zahn, *Florida State University* • P. V. Rao, *University of Florida* • Laura Reisert, *Florida International University* • Jeffrey W. Steagall, *University of North Florida* • Edna White, *Florida Atlantic University* **GEORGIA** Robert Elrod, *Georgia State University* • Karen Smith, *West Georgia University* **HAWAII** Steve Hora, *University of Hawaii, Hilo* **ILLINOIS** Arunas Dagys, *St. Xavier University* • Edward Minieka, *University of Illinois at Chicago* • Don Robinson, *Illinois State University* • Chipei Tseng, *Northern Illinois University* • Pankaj Vaish, *Arthur Andersen & Company* **IOWA** Dileep Dhavale, *University of Northern Iowa* • William Duckworth II, William Q. Meeker, *Iowa State University* • Tim E. McDaniel, *Buena Vista University* **KANSAS** Paul I. Nelson, *Kansas State University* • Lawrence A. Sherr, *University of Kansas* **KENTUCKY** Richard N. McGrath, *Bowling Green State University* **LOUISIANA** James Willis, *Louisiana State University* **MARYLAND** John F. Beyers, Michael Kulansky, *University of Maryland–University College* • Glenn J. Browne, Mary C. Christman, *University of Maryland* **MASSACHUSETTS** Warren M. Holt, *Southeastern Massachusetts University* • Remus Osan, *Boston University* **MICHIGAN** Atul Agarwal, Petros Ioannatos, *GMI Engineering and Management Institute* • Richard W. Andrews, Peter Lenk, Benjamin Lev, *University of Michigan* • Leszek Gawarecki, *Kettering University* • Toni M. Somers, *Wayne State University* • William Welch, *Saginaw Valley State University* • T. J. Wharton, *Oakland University* **MINNESOTA** Gordon J. Alexander, Donald W. Bartlett, David M. Bergman, Atul Bhatia, Steve Huchendorf, Benny Lo, Karen Lundquist, Vijay Pisharody, Donald N. Steinnes, Robert W. Van Cleave, Steve Wickstrom, *University*

of Minnesota • Daniel G. Brick, Leigh Lawton, *University of St. Thomas* • Susan Flach, *General Mills, Inc.* • David D. Krueger, Ruth K. Meyer, Jan Saraph, Gary Yoshimoto, *St. Cloud State University* • Paula M. Oas, *General Office Products* • Fike Zahroom, *Moorhead State University* **MISSISSIPPI** Eddie M. Lewis, *University of Southern Mississippi* • Alireza Tahai, *Mississippi State University* **MISSOURI** James Holstein, Lawrence D. Ries, *University of Missouri, Columbia* • Marius Janson, L. Douglas Smith, *University of Missouri, St. Louis* • Farroll Tim Wright, *University of Missouri* • Stephanie Schartel-Dunn *Missouri Western University* **NEBRASKA** James Wright, *Chadron State College* **NEW HAMPSHIRE** Ken Constantine, *University of New Hampshire* **NEW JERSEY** Lewis Coopersmith, Cengiz Haksever, *Rider University* • Lei Lei, Xuan Li, Zina Taran, *Rutgers University* • Philip Levine, Leonard Presby, *William Paterson University* **NEW MEXICO** S. Howard Kraye, *University of New Mexico* **NEW YORK** James Czachor, *Fordham-Lincoln Center, AT&T* • Bernard Dickman, *Hofstra University* • Joshua Fogel, *Brooklyn College of City University of New York* • Martin Labbe, *State University of New York, College at New Paltz* • Kenneth Leong, *College of New Rochelle* • Mark R. Marino, *Niagara University/Erie Community College* • G. E. Martin, *Clarkson University* • Thomas J. Pfaff, *Ithaca College* • Gary Simon, *New York University, Stern School of Business* • Rungrudee Suetorsak, *SUNY-Fredonia* **NORTH CAROLINA** Golam Azam, *North Carolina Agricultural & Technical University* • Edward Carlstein, Douglas A. Elvers, *University of North Carolina at Chapel Hill* • Barry P. Cuffe, *Wingate University* • Don Holbert, *East Carolina University* • J. Morgan Jones, *University of North Carolina* • Douglas S. Shafer, *University of North Carolina, Charlotte* **OHIO** William H. Beyer, *University of Akron* • Michael Broida, Tim Krehbiel, *Miami University of Ohio* • Chih-Hsu Cheng, Douglas A. Wolfe, *Ohio State University* • Ronald L. Coccari, *Cleveland State University* • Richard W. Culp, *Wright-Patterson AFB, Air Force Institute of Technology* **OKLAHOMA** Larry Claypool, Brenda Masters, Rebecca Moore, *Oklahoma State University* • Robert Curley, *University of Central Oklahoma* **PENNSYLVANIA** Mohammed Albohali, Douglas H. Frank, *Indiana University of Pennsylvania* • Sukhwinder Bagi, *Bloomsburg University* • Carl Bedell, *Philadelphia College of Textiles and Science* • Ann Hussein, *Philadelphia University* • Behnam Nakhai, *Millersville University* • Rose Prave, *University of Scranton* • Farhad Saboori, *Albright College* • Kathryn Szabet, *LaSalle University* • Pandu Tadikamalla *University of Pittsburgh* • Christopher J. Zappe, *Bucknell University* **SOUTH CAROLINA** Iris Fetta, Robert Ling, *Clemson University* • Kathleen M. Whitcomb, *University of South Carolina* **TENNESSEE** Francis J. Brewerton, *Middle Tennessee State University* **TEXAS** Larry M. Austin, *Texas Tech University* • Jim Branscome, Robert W. Brobst, Mark Eakin, Grace Esimai, Michael E. Hanna, Craig W. Slinkman, *University of Texas at Arlington* • Katarina Jegdic, *University of Houston–Downtown* • Virgil F. Stone, *Texas A & M University* **VIRGINIA** Edward R. Clayton, *Virginia Polytechnic Institute and State University* **WASHINGTON** June Morita, Kim Tamura, *University of Washington* **WISCONSIN** Ross H. Johnson, *Madison College* **WASHINGTON, D.C.** Keith Ord, *Georgetown University* • Balaji Srinivasan, *George Washington University* **CANADA** Clarence Bayne, *Concordia University* • Edith Gombay, *University of Alberta* **TURKEY** Dilek Onkal, *Bilkent University, Ankara* **OTHER** Michael P. Wegmann, *Keller Graduate School of Management*

Other Contributors

Special thanks are due to our supplements author, Mark Dummeldinger, who has worked with us for many years. Accuracy checker Engin Sungur helped ensure a highly accurate, clean text. Nati Jain and Shantel Vargas diligently researched new data and cases for the exercise sets. Finally, the Pearson Education staff of Amanda Brands, Suzanna Bainbridge, Karen Montgomery, Alicia Wilson, Demetrius Hall, Joe Vetere, Peggy McMahon, Jean Choe, and Bob Carroll, who helped greatly with all phases of the text development, production, and marketing effort. Our gratitude also goes to Heidi Aguiar of Spi Global for overseeing the production process.

MyLab Statistics Resources for Success

MyLab Statistics is available to accompany Pearson's market-leading text options, including Statistics for Business and Economics, 14th Edition (access code required).

MyLab™ is the teaching and learning platform that empowers you to reach every student. MyLab Statistics combines trusted author content—including full eText and assessment with immediate feedback—with digital tools and a flexible platform to personalize the learning experience and improve results for each student. Integrated with StatCrunch®, an web-based statistical software program, students learn the skills they need to interact with data in the real world.

Student Resources

Student's Solutions Manual, by Mark Dummeldinger (University of South Florida), provides detailed, worked-out solutions to all odd-numbered text exercises. (ISBN-10: 0-13-685540-7, ISBN-13: 978-0-13-685540-8)

Excel Technology Manual, by Mark Dummeldinger (University of South Florida), provides tutorial instruction and worked-out text examples for Excel. The *Excel Technology Manual* is available for download at pearsonhighered.com/mathstatresrouces or within MyLab Statistics.

NEW! Videos by author Terry Sincich. These videos illustrate the solutions to select exercises in the book, demonstrate key topics in detail, and include tutorials on real-world statistics in action.

Accessibility

Pearson works continuously to ensure our products are as accessible as possible to all students. Currently we work toward achieving WCAG 2.0 AA for our existing products (2.1 AA for future products) and Section 508 standards, as expressed in the Pearson Guidelines for Accessible Educational Web Media.

For the Instructor

Annotated Instructor's Edition contains answers to text exercises. Annotated marginal notes include Teaching Tips, suggested exercises to reinforce the statistical concepts discussed in the text, and short answers to the exercises within the exercise sets. (ISBN-10: 0-13-685542-3; ISBN-13: 978-0-13-685542-2)

Instructor's Solutions Manual, by Mark Dummeldinger (University of South Florida), provides detailed, worked-out solutions to all of the book's exercises. Careful attention has been paid to ensure that all methods of solution and notation are consistent with those used in the core text. Available for download at www.pearson.com.

TestGen® (www.pearsoned.com/testgen) enables instructors to build, edit, print, and administer tests using a computerized bank of questions developed to cover all the objectives of the text. TestGen is algorithmically based, allowing instructors to create multiple but equivalent versions of the same question or test with the click of a button. Instructors can also modify test bank questions or add new questions. The software and testbank are available for download from Pearson's Instructor Resource Center.

Data Analytics/Early Alerts

Instructors have a comprehensive gradebook with enhanced reporting functionality that makes it easier to understand which students are struggling, and which topics they struggle with most. New Early Alerts use predictive analytics to identify struggling students as early as possible—even if their assignment scores are not a cause for concern.

pearson.com/mylab/statistics

1

CONTENTS

WHERE WE'RE GOING

- Introduce the field of statistics **(1.1)**
- Demonstrate how statistics applies to business **(1.2)**
- Introduce the language of statistics and the key elements of any statistical problem **(1.3)**
- Differentiate between population and sample data **(1.3)**
- Differentiate between descriptive and inferential statistics **(1.3)**
- Introduce the key elements of a process **(1.4)**
- Identify the different types of data and data-collection methods **(1.5–1.6)**
- Discover how critical thinking through statistics can help improve our quantitative literacy **(1.7)**

Statistics, Data, and Statistical Thinking

STATISTICS IN ACTION — A 20/20 View of Surveys and Studies: Facts or Fake News?

Several years ago, the popular ABC television program 20/20 aired a story titled "Fact or Fiction?—Exposés of So-Called Surveys." The focus of the program segment was on what we now call "fake news" or "alternative facts," that is, false information that is often highly publicized in the media (newspapers, magazines, TV shows, Twitter, Instagram, etc.). Several misleading (and possibly unethical) surveys, conducted by businesses or special interest groups with specific objectives in mind, were presented on the ABC program. Several are listed in Table SIA1.1, as well as some recent misleading studies used in product advertisements.

For this *Statistics in Action*, we consider research sponsored by Mars Corp. and published in the journal *Nature Neuroscience* (Dec. 2014). Researchers from Columbia University wondered whether taking cocoa supplements would enhance a region of the brain that deteriorates with age and is associated with age-related memory loss. They concluded that cocoa supplements can indeed boost cognition in older adults. Other similar studies claim that chocolate will reduce cardiovascular disease risk and help with weight loss. These results were reported on by media outlets such as the *New York Times*, with headlines like, "*To improve a memory, consider chocolate!*", "*Good news for chocolate lovers: The more you eat, the lower your risk of heart disease*," or simply "*Chocolate is good for you*." These reported "facts" have likely grown consumer demand for chocolate. At a time when candy sales overall has declined, chocolate retail sales in the United States have risen from $14.2 billion in 2007 to $18.9 billion in 2017.

Critical thinkers may question chocolate as a powerful health food. One such group, Vox Media (November 7, 2017), investigated this chocolate phenomenon.

STATISTICS IN ACTION (*continued*)

First, Vox discovered that chocolate companies like Mars, "have spent millions of dollars for scientific studies and research grants that support cocoa science. And, of 100 [recent] Mars-sponsored studies on cocoa, chocolate, and health, 98 had conclusions that were favorable to the candy maker in some way." This unusually high percentage of favorable studies, according to Vox, "raises questions about the quality of the studies, given that Mars and other chocolate makers can use the positive findings to market their products." This finding motivated Vox Media to critically analyze the *Nature Neuroscience* chocolate study.

To conduct this study, the Vox Media researchers randomly assigned 37 people to one of four groups. Each subject in group 1 was given a high daily dose (900 mg) of cocoa flavanol supplements and assigned one hour of aerobic exercise four times per week. Subjects in group 2 received the same high dose of cocoa flavanol supplements but were not assigned to exercise. Group 3 subjects received a low dose of cocoa flavanols (10 mg) and were assigned one hour of aerobic exercise four times per week. Finally, the last group received a low dose of cocoa flavanols but was not assigned exercise (See Figure SIA1.1). After a 3-month period, the researchers tested whether cocoa flavanol supplements staved off cognitive decline in a region of the brain associated with age-related memory loss. They did this by measuring brain waves in an MRI machine and by using an object-recognition task to test memory and reaction time. The researchers also tested if exercise had any effect on memory.

Group 1: High Cocoa / Exercise (8 subjects)	*Group 2*: High Cocoa / No Exercise (11 subjects)
Group 3: Low Cocoa / Exercise (9 subjects)	*Group 4*: Low Cocoa / No Exercise (9 subjects)

Figure SIA1.1
Schematic of Chocolate Study

The researchers reported that exercise had no impact on brain function—but cocoa flavanols did. Subjects receiving a high cocoa flavanol dosage demonstrated a greater improvement in cognitive performance than those in the low dosage groups. However, as reported by Vox Media, the researchers drew conclusions that went beyond the scope of the study. For example, the researchers claimed that the effects they saw in the high-flavanol group demonstrated that cocoa could reverse age-related memory decline by 30 years. Vox also discovered problems with the study's small sample size and design.

Henry Drysdale, a doctor and fellow at Oxford University's Center for Evidence-Based Medicine elaborated on the design issue. First, he warned that eating cocoa supplements in order to improve memory in three months is not relevant to real-world age-related memory decline. Second, the doctor pointed out the need for a much larger (than 37) group of study participants, and to conduct the trial for several years. Finally, Drysdale commented on the study variables: "Instead of only tracking the study participants' brain waves in an MRI machine (which is not a measure of cognitive ability), or using an object recognition task to test memory, you'd also want to measure outcomes that matter in people's lives, like, whether those taking cocoa could remember what they did that morning or that they had a doctor's appointment next week better than the people who didn't take the cocoa."

Ultimately, Vox Media stated that, "this trial only demonstrated that supplements seem to enhance brain function over a period of weeks, and only according to a very specific (and not very widely used) test of cognitive function. That is far from valid proof that cocoa is a memory enhancer."

To conclude the introduction to this *Statistics in Action*, we return to the ABC *20/20* TV program segment. The segment ended with an interview of Cynthia Crossen, author of *Tainted Truth: The Manipulation of Fact in America*, an exposé of misleading and biased surveys.

Some 20 years before the term "fake news" was coined, Crossen warned, "If everybody is misusing numbers and scaring us with numbers to get us to do something, however good [that something] is, we've lost the power of numbers. Now, we know certain things from research. For example, we know that smoking cigarettes is hard on your lungs and heart, and because we know that, many people's lives have been extended or saved. We don't want to lose the power of information to help us make decisions, and that's what I worry about."

Table SIA1.1: Examples of "Fake News"

Fake News (Source)	Actual Study Information/Flaw
1. Eating oat bran is a cheap and easy way to reduce your cholesterol. (*Quaker Oats*)	Diet must consist of nothing but oat bran to reduce your cholesterol count.
2. One in four American children under age 12 is hungry or at risk of hunger. (*Food Research and Action Center*)	Based on responses to questions: "Do you ever cut the size of meals?" "Do you ever eat less than you feel you should?" "Did you ever rely on limited numbers of foods to feed your children because you were running out of money to buy food for a meal?"
3. There is a strong correlation between a CEO's golf handicap and the company's stock performance: The lower the CEO's handicap (i.e., the better the golfer), the better the stock performs. (*New York Times*, May 31, 1998)	Survey sent to CEOs of 300 largest US companies; only 51 revealed their golf handicaps. Data for several top-ranking CEOs were excluded from the analysis.
4. Prior to the passing of the federal government's health reform act, 30% of employers are predicted to "definitely" or "probably" stop offering health coverage. (*McKinsey & Company Survey*, February 2011)	Online survey of 1,329 private-sector employers in the United States. Respondents were asked leading questions that made it logical to stop offering health insurance.
5. In an advertisement, "more than 80% of dentists surveyed recommend Colgate toothpaste to patients." (*Colgate-Palmolive Company*, January 2007)	The survey allowed each dentist to recommend more than one toothpaste.The Advertising Standards Authority cited and fined Colgate for a misleading ad (implying 80% of dentists recommend Colgate toothpaste in preference to all other brands) and banned the advertisement.
6. An advertisement for Kellogg's Frosted Mini-Wheats claimed that the cereal was "clinically shown to improve kids' attentiveness by nearly 20%." (*Kellogg Company*, 2009)	Only half of the kids in the study showed any improvement in attentiveness; only 1 in 7 improved by 18% or more, and only 1 in 9 improved by 20% or more; kids who ate Frosted Mini-Wheats were compared against kids who had only water for breakfast. (The Kellogg Company agreed to pay $4 million to settle suit over false ad claim.)
7. On the basis of a commissioned study, Walmart advertised that it "was responsible for an overall 3.1% decline in consumer prices" and it "saves customers over $700 per year." (*Global Insight*, 2005)	The Economic Policy Institute noted that the Global Insight study was based on the retailer's impact on the Consumer Price Index (CPI)—but 60% of the items in the CPI are services, not commodities that can be purchased at Walmart. (Walmart was forced to withdraw the misleading advertisement.)
8. In a survey commissioned by cable provider Comcast, respondents were asked to decide which cable provider, Comcast or DIRECTV, offered more HD channels. Respondents were shown channel lists for DIRECTV (List #387) and Comcast (List #429). (*NAD Case Report No. 5208*, August 25, 2010).	The National Advertising Division (NAD) of the Council of Better Business Bureaus rejected the survey after finding that the higher list number (#429) "served as a subtle, yet effective cue" that Comcast's list contained more channels.

STATISTICS IN ACTION (*continued*)

Fake News (Source)	Actual Study Information/Flaw
9. NPR reported on a recent study that found that teens who spend five or more hours per day on their smartphones are 71 percent more likely to have one risk factor for suicide/depression (*Journal of Abnormal Psychology*, March 2019).	Data was based on teen's perceived use of their smartphones. Studies have shown that perceived use is poorly related to actual use measured with an app. Also, the measures of addiction or suicide risk were based on unreliable scales.

In the following *Statistics in Action Revisited* sections, we discuss several key statistical concepts covered in this chapter that are relevant to misleading surveys and studies.

STATISTICS IN ACTION REVISITED

- Identifying the population, sample, and inference **(p. 11)**
- Identifying the data-collection method and data type **(p. 21)**
- Critically assessing the ethics of a statistical study **(p. 24)**

1.1 The Science of Statistics

What does *statistics* mean to you? Does it bring to mind batting averages? Gallup polls, unemployment figures, or numerical distortions of facts (lying with statistics)? Or is it simply a college requirement you have to complete? We hope to persuade you that statistics is a meaningful, useful science with a broad scope of applications to business, government, and the physical and social sciences that is almost limitless. We also want to show that statistics can lie only when they are misapplied. Finally, we wish to demonstrate the key role statistics play in critical thinking—whether in the classroom, on the job, or in everyday life. Our objective is to leave you with the impression that the time you spend studying this subject will repay you in many ways.

Although the term can be defined in many ways, a broad definition of *statistics* is the science of collecting, classifying, analyzing, and interpreting information. Thus, a statistician isn't just someone who calculates batting averages at baseball games or tabulates the results of a Gallup poll. Professional statisticians are trained in *statistical science*—that is, they are trained in collecting information in the form of **data,** evaluating it, and drawing conclusions from it. Furthermore, statisticians determine what information is relevant in a given problem and whether the conclusions drawn from a study are to be trusted.

> **Statistics** is the science of data. It involves collecting, classifying, summarizing, organizing, analyzing, and interpreting numerical and categorical information.

In the next section, you'll see several real-life examples of statistical applications in business and government that involve making decisions and drawing conclusions.

1.2 Types of Statistical Applications in Business

Statistics means "numerical descriptions" to most people. Monthly unemployment figures, the failure rate of startup companies, and the proportion of female executives in a particular industry all represent statistical descriptions of large sets of data collected on some phenomenon. Often the data are selected from some larger set of data that has characteristics we wish to estimate. We call this selection process *sampling*. For example, you might collect the ages of a sample of customers of a video streaming services company to estimate the average age of *all* customers of the company. Then you could use your estimate to target the firm's advertisements to the appropriate age group. Notice that statistics involves two different processes: (1) describing sets of data and (2) drawing conclusions (making estimates, decisions, predictions, etc.) about the sets of data based on sampling. So, the applications of statistics can be divided into two broad areas: *descriptive statistics* and *inferential statistics*.

BIOGRAPHY

FLORENCE NIGHTINGALE (1820–1910)
The Passionate Statistician

In Victorian England, the "Lady of the Lamp" had a mission to improve the squalid field hospital conditions of the British army during the Crimean War. Today, most historians consider Florence Nightingale to be the founder of the nursing profession. To convince members of the British Parliament of the need for supplying nursing and medical care to soldiers in the field, Nightingale compiled massive amounts of data from the army files. Through a remarkable series of graphs (which included the first "pie chart"), she demonstrated that most of the deaths in the war were due to illnesses contracted outside the battlefield or long after battle action from wounds that went untreated. Florence Nightingale's compassion and self-sacrificing nature, coupled with her ability to collect, arrange, and present large amounts of data, led some to call her the "Passionate Statistician."

Descriptive statistics utilizes numerical and graphical methods to explore data, i.e., to look for patterns in a data set, to summarize the information revealed in a data set, and to present the information in a convenient form for the user.

Inferential statistics utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data.

Although we'll discuss both descriptive and inferential statistics in the following chapters, the primary theme of the text is **inference.**

Let's begin by examining some business studies that illustrate applications of statistics.

Study 1.1 "Best-Selling Girl Scout Cookies" (Source: www.girlscouts.org): Since 1917, the Girl Scouts of America have been selling boxes of cookies. In 2017, there were 12 varieties for sale: Thin Mints, Samoas, Lemonades, Tagalongs, Do-si-dos, Trefoils, Savannah Smiles, Thanks-A-Lot, Dulce de Leche, Cranberry Citrus Crisps, Chocolate Chip, and Thank U Berry Much. Each of the approximately 150 million boxes of Girl Scout cookies sold in 2017 was classified by variety. The results are summarized in Figure 1.2. From the graph, you can clearly see that the best-selling variety is Thin Mints (25%), followed by Samoas (19%) and Tagalongs (13%). Since Figure 1.1 *describes* the variety of categories of the boxes of Girl Scout cookies sold, the graphic is an example of *descriptive statistics.*

Study 1.2 "Executive Compensation vs. Typical Worker Pay" (*Source*: 24/7 Wall Street, *USA Today*, October 15, 2018): How big is the gap between what a firm pays its CEO and what it pays its typical worker? To answer this question, 24/7 Wall Street reviewed the ratio between annual CEO base pay and typical worker salary at 168 large US companies, using data from benefits and compensation information provided by the website Payscale. This information was used to compute the ratio of CEO pay to the typical worker salary

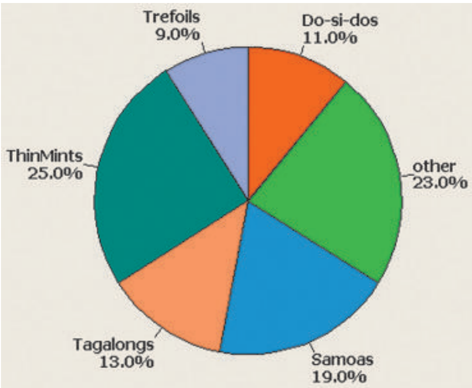


Figure 1.1
Best-selling Girl Scout cookies
Source: "Best-Selling Girl Scout Cookies," based on www.girlscouts.org.

at each company.* The data for the 10 companies with the highest ratio in the sample of 168 companies in the study are shown in Table 1.1. An analysis of the data for all 168 firms revealed that the "average" ratio of CEO pay to typical worker pay was 205.† In other words, on average, CEOs in the sample earn around 205 times what their firm's typical worker earns. Armed with this sample information, an economist might *infer* that the average ratio of CEO pay to typical worker pay for *all* US firms is 205. Thus, this study is an example of *inferential statistics*.

Table 1.1 Ratio of CEO Compensation to Typical Worker Pay

Company	CEO	CEO Base Pay	CEO Total Compensation	Typical Worker Pay	Ratio
1 **CVS Health**	Larry J. Merlo	$12,105,481	$22,855,374	$27,900	434
2 **CBS Corp.**	Leslie Moonves	$23,652,883	$56,352,801	$59,900	395
3 **Walt Disney**	Robert A. Iger	$26,208,003	$43,490,567	$71,400	367
4 **TGX Comp.**	Carol Meyrowitz	$7,330,584	$17,962,232	$22,400	327
5 **21st Century Fox**	K. Rupert Murdoch	$17,047,636	$22,192,923	$54,800	311
6 **Comcast**	Brian L. Roberts	$16,819,942	$27,520,744	$55,800	301
7 **L Brands**	Leslie H. Wexner	$9,665,925	$26,669,306	$33,900	285
8 **Honeywell Int.**	David M. Cote	$22,767,851	$33,105,851	$81,600	279
9 **PepsiCo**	Indra K. Nooyi	15,937,828	$22,189,307	$61,500	259
10 **Wynn Resorts**	Stephen A. Wynn	$11,930,391	$20,680,391	$50,100	238

Source: 24/7 Wall Street, *USA Today*, Oct. 15, 2018.

Study 1.3 "Does rudeness really matter in the workplace?" (*Academy of Management Journal*, October 2007): Many studies have established that rudeness in the workplace can lead to retaliatory and counterproductive behavior. However, there has been little research on how rude behaviors influence a victim's task performance. In one study, college students enrolled in a management course were randomly assigned to one of two experimental conditions: rudeness condition (45 students) and control group (53 students). Each student was asked to write down as many uses for a brick as possible in 5 minutes; this value (total number of uses) was used as a performance measure for each student. For those students in the rudeness condition, the facilitator displayed rudeness by berating the students in general for being irresponsible and unprofessional (due to a late-arriving associate of the researchers). No comments were made about the late-arriving associate of the researchers to students in the control group. As you might expect, the researchers discovered that the performance levels for students in the rudeness condition were generally lower than the performance levels for students in the control group; thus, they concluded that rudeness in the workplace negatively affects job performance. As in Study 1.2, this study is an example of the use of inferential statistics. The researchers used data collected on 98 college students in a simulated work environment to make an inference about the performance levels of all workers exposed to rudeness on the job.

These studies provide three real-life examples of the uses of statistics in business, economics, and management. Notice that each involves an analysis of data, either for the purpose of describing the data set (Study 1.1) or for making inferences about a data set (Studies 1.2 and 1.3).

*The ratio was calculated using the *median* worker salary at each firm. A formal definition of median is given in Chapter 2. For now, think of the median as the *typical* salary for a worker, i.e., one that falls in the middle of all worker salaries.

†A formal definition of *average* is also given in Chapter 2. Like the median, think of the average as another way to express the *middle* salary.

1.3 Fundamental Elements of Statistics

Statistical methods are particularly useful for studying, analyzing, and learning about *populations* of *experimental units*.

An **experimental (or observational) unit** is an object (e.g., person, thing, transaction, or event) upon which we collect data.

A **population** is a set of units (usually people, objects, transactions, or events) that we are interested in studying.

For example, populations may include (1) *all* employed workers in the United States; (2) *all* registered voters in California; (3) *everyone* who has purchased a particular brand of cell phone; (4) *all* the cars produced last year by a particular assembly line; (5) the *entire* stock of spare parts at United Airlines' maintenance facility; (6) *all* sales made at the drive-through window of a McDonald's restaurant during a given year; and (7) the set of *all* accidents occurring on a particular stretch of interstate during a holiday period. Notice that the first three population examples (1–3) are sets (groups) of people, the next two (4–5) are sets of objects, the next (6) is a set of transactions, and the last (7) is a set of events. Also notice that *each set includes all the experimental units in the population* of interest.

In studying a population, we focus on one or more characteristics or properties of the experimental units in the population. We call such characteristics *variables*. For example, we may be interested in the variables age, gender, income, and/or the number of years of education of the people currently unemployed in the United States.

A **variable** is a characteristic or property of an individual experimental (or observational) unit.

The name *variable* is derived from the fact that any particular characteristic may vary among the experimental units in a population.

In studying a particular variable, it is helpful to be able to obtain a numerical representation for it. Often, however, numerical representations are not readily available, so the process of measurement plays an important supporting role in statistical studies. **Measurement** is the process we use to assign numbers to variables of individual population units. We might, for instance, measure the preference for a food product by asking a consumer to rate the product's taste on a scale from 1 to 10. Or we might measure workforce age by simply asking each worker, "How old are you?" In other cases, measurement involves the use of instruments such as timers, scales, and calipers.

If the population we wish to study is small, it is possible to measure a variable for every unit in the population. For example, if you are measuring the starting salary for all University of Michigan MBA graduates last year, it is at least feasible to obtain every salary. When we measure a variable for every experimental unit of a population, the result is called a **census** of the population. Typically, however, the populations of interest in most applications are much larger, involving perhaps many thousands or even an infinite number of units. Examples of large populations include the seven listed above, as well as all invoices produced in the last year by a *Fortune* 500 company, all potential buyers of a new iPad, and all stockholders of a firm listed on the New York

Stock Exchange. For such populations, conducting a census would be prohibitively time-consuming and/or costly. A reasonable alternative would be to select and study a *subset* (or portion) of the units in the population.

A **sample** is a subset of the units of a population.

For example, suppose a company is being audited for invoice errors. Instead of examining all 15,472 invoices produced by the company during a given year, an auditor may select and examine a sample of just 100 invoices (see Figure 1.2). If he is interested in the variable "invoice error status," he would record (measure) the status (error or no error) of each sampled invoice.

After the variable(s) of interest for every experimental unit in the sample (or population) is (are) measured, the data are analyzed, either by descriptive or by inferential statistical methods. The auditor, for example, may be interested only in *describing* the error rate in the sample of 100 invoices. More likely, however, he will want to use the information in the sample to make *inferences* about the population of all 15,472 invoices.

A **statistical inference** is an estimate or prediction or some other generalization about a population based on information contained in a sample.

*That is, we use the information contained in the sample to learn about the larger population.** Thus, from the sample of 100 invoices, the auditor may estimate the total number of invoices containing errors in the population of 15,472 invoices. The auditor's inference about the quality of the firm's invoices can be used in deciding whether to modify the firm's billing operations.

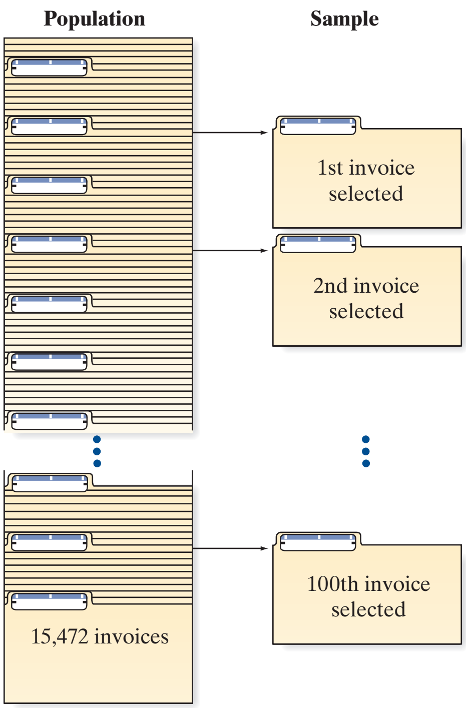


Figure 1.2
A sample of all company invoices

*The terms *population* and *sample* are often used to refer to the sets of measurements themselves, as well as to the units on which the measurements are made. When a single variable of interest is being measured, this usage causes little confusion. But when the terminology is ambiguous, we'll refer to the measurements as *population data sets* and *sample data sets*, respectively.

EXAMPLE 1.1

Key Elements of a Statistical Problem—Ages of Cable TV News Viewers

Problem According to the most recent Nielsen survey of cable TV news viewers, the average age of CNN viewers is 60 years. Suppose a rival network (e.g., FOX) executive hypothesizes that the average age of FOX viewers is greater than 60. To test her hypothesis, she samples 200 FOX viewers and determines the age of each.

a. Describe the population.
b. Describe the variable of interest.
c. Describe the sample.
d. Describe the inference.

Solution

a. The population is the set of units of interest to the TV executive, which is the set of all FOX viewers.
b. The age (in years) of each viewer is the variable of interest.
c. The sample must be a subset of the population. In this case, it is the 200 FOX viewers selected by the executive.
d. The inference of interest involves the *generalization* of the information contained in the sample of 200 viewers to the population of all FOX viewers. In particular, the executive wants to *estimate* the average age of the viewers in order to determine whether it exceeds 60 years. She might accomplish this by calculating the average age in the sample and using the sample average to estimate the population average.

Look Back A key to diagnosing a statistical problem is to identify the data set collected (in this example, the ages of the 200 FOX TV viewers) as a population or sample.

EXAMPLE 1.2

Key Elements of a Statistical Problem—Pepsi vs. Coca-Cola

Problem *Cola wars* is the popular term for the intense competition between Coca-Cola and Pepsi displayed in their marketing campaigns. Their campaigns have featured claims of consumer preference based on taste tests. For example, the *Huffington Post* (November 11, 2013) conducted a blind taste test of 9 cola brands that included Coca-Cola and Pepsi. (Pepsi finished 1st and Coke finished 5th.) Suppose, as part of a Pepsi marketing campaign, 1,000 cola consumers are given a blind taste test (i.e., a taste test in which the two brand names are disguised). Each consumer is asked to state a preference for brand A or brand B.

a. Describe the population.
b. Describe the variable of interest.
c. Describe the sample.
d. Describe the inference.

Solution

a. Because we are interested in the responses of cola consumers in a taste test, a cola consumer is the experimental unit. Thus, the population of interest is the collection or set of all cola consumers.
b. The characteristic that Pepsi wants to measure is the consumer's cola preference as revealed under the conditions of a blind taste test, so cola preference is the variable of interest.
c. The sample is the 1,000 cola consumers selected from the population of all cola consumers.
d. The inference of interest is the *generalization* of the cola preferences of the 1,000 sampled consumers to the population of all cola consumers. In particular, the

preferences of the consumers in the sample can be used to *estimate* the percentage of all cola consumers who prefer each brand.

Look Back In determining whether the statistical application is inferential or descriptive, we assess whether Pepsi is interested in the responses of only the 1,000 sampled customers (descriptive statistics) or in the responses for the entire population of consumers (inferential statistics).

Now Work Exercise 1.16

The preceding definitions and examples identify four of the five elements of an inferential statistical problem: a population, one or more variables of interest in a sample, and an inference. But making the inference is only part of the story. We also need to know its **reliability**–that is, how good the inference is. The only way we can be certain that an inference about a population is correct is to include the entire population in our sample. However, because of *resource constraints* (e.g., insufficient time and/or money), we usually can't work with whole populations, so we base our inferences on just a portion of the population (a sample). Consequently, whenever possible, it is important to determine and report the reliability of each inference made. Reliability, then, is the fifth element of inferential statistical problems.

The measure of reliability that accompanies an inference separates the science of statistics from the art of fortune-telling. A palm reader, like a statistician, may examine a sample (your hand) and make inferences about the population (your life). However, unlike statistical inferences, the palm reader's inferences include no measure of reliability.

Suppose, like the TV executive in Example 1.1, we are interested in the *error of estimation* (i.e., the difference between the average age of the population of TV viewers and the average age of a sample of TV viewers). Using statistical methods, we can determine a *bound on the estimation error*. This bound is simply a number that our estimation error (the difference between the average age of the sample and the average age of the population) is not likely to exceed. We'll see in later chapters that bound is a measure of the uncertainty of our inference. The reliability of statistical inferences is discussed throughout this text. For now, we simply want you to realize that an inference is incomplete without a measure of its reliability.

A **measure of reliability** is a statement (usually quantified) about the degree of uncertainty associated with a statistical inference.

Let's conclude this section with a summary of the elements of both descriptive and inferential statistical problems and an example to illustrate a measure of reliability.

Four Elements of Descriptive Statistical Problems

1. The population or sample of interest
2. One or more variables (characteristics of the population or experimental units) that are to be investigated
3. Tables, graphs, or numerical summary tools
4. Identification of patterns in the data

Five Elements of Inferential Statistical Problems

1. The population of interest
2. One or more variables (characteristics of the population or experimental units) that are to be investigated
3. The sample of population units
4. The inference about the population based on information contained in the sample
5. A measure of reliability for the inference

EXAMPLE 1.3

Reliability of an Inference—Pepsi vs. Coca-Cola

Problem Refer to Example 1.2, in which the cola preferences of 1,000 consumers were indicated in a taste test. Describe how the reliability of an inference concerning the preferences of all cola consumers in the Pepsi bottler's marketing region could be measured.

Solution When the preferences of 1,000 consumers are used to estimate the preferences of all consumers in the region, the estimate will not exactly mirror the preferences of the population. For example, if the taste test shows that 56% of the 1,000 consumers chose Pepsi, it does not follow (nor is it likely) that exactly 56% of all cola drinkers in the region prefer Pepsi. Nevertheless, we can use sound statistical reasoning (which is presented later in the text) to ensure that our sampling procedure will generate estimates that are almost certainly within a specified limit of the true percentage of all consumers who prefer Pepsi. For example, such reasoning might assure us that the estimate of the preference for Pepsi from the sample is almost certainly within 5% of the actual population preference. The implication is that the actual preference for Pepsi is between 51% [i.e., $(56 - 5)\%$] and 61% [i.e., $(56 + 5)\%$]—that is, $(56 \pm 5)\%$. This interval represents a measure of reliability for the inference.

Look Back The interval 56 ± 5 is called a *confidence interval*, because we are "confident" that the true percentage of customers who prefer Pepsi in a taste test falls into the range (51, 61). In Chapter 6, we learn how to assess the degree of confidence (e.g., 90% or 95% confidence) in the interval.

STATISTICS IN ACTION REVISITED

Identifying the Population, Sample, and Inference

Consider the study on the link between a CEO's golf handicap and the company's stock performance, reported in the *New York Times*. The newspaper gathered information on golf handicaps of corporate executives obtained from a *Golf Digest* survey sent to CEOs of the 300 largest US companies. (A golf handicap is a numerical "index" that allows golfers to compare skills; the lower the handicap, the better the golfer.) For the 51 CEOs who reported their handicaps, the *New York Times* then determined each CEO's company stock market performance over a 3-year period (measured as a rate-of-return index, from a low value of 0 to a high value of 100). Thus, the experimental unit for the study is a corporate executive, and the two variables measured are golf handicap and stock performance index. Also, the data for the 51 CEOs represent a sample selected from the much larger population of all corporate executives in the United States. (These data are available in the **GLFCEO** file.)

The *New York Times* discovered a "statistical correlation" (a method discussed in Chapter 11) between golf handicap and stock performance. Thus, the newspaper inferred that the better the CEO is at golf, the better the company's stock performance.

Data Set: GLFCEO

1.4 Processes (Optional)

Sections 1.2 and 1.3 focused on the use of statistical methods to analyze and learn about populations, which are sets of *existing* units. Statistical methods are equally useful for analyzing and making inferences about *processes*.

A **process** is a series of actions or operations that transforms inputs to outputs. A process produces or generates output over time.

The most obvious processes of interest to businesses are those of production or manufacturing. A manufacturing process uses a series of operations performed by

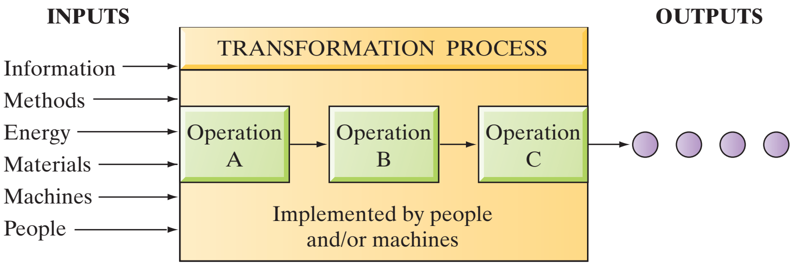


Figure 1.3
Graphical depiction of a manufacturing process

people and machines to convert inputs, such as raw materials and parts, to finished products (the outputs). Examples include the process used to produce the paper on which these words are printed, automobile assembly lines, and oil refineries.

Figure 1.3 presents a general description of a process and its inputs and outputs. In the context of manufacturing, the process in the figure (i.e., the transformation process) could be a depiction of the overall production process or it could be a depiction of one of the many processes (sometimes called *subprocesses*) that exist within an overall production process. Thus, the output shown could be finished goods that will be shipped to an external customer or merely the output of one of the steps or subprocesses of the overall process. In the latter case, the output becomes input for the next subprocess. For example, Figure 1.4 could represent the overall automobile assembly process, with its output being fully assembled cars ready for shipment to dealers. Or, it could depict the windshield assembly subprocess, with its output of partially assembled cars with windshields ready for "shipment" to the next subprocess in the assembly line.

Besides physical products and services, businesses and other organizations generate streams of numerical data over time that are used to evaluate the performance of the organization. Examples include weekly sales figures, quarterly earnings, and yearly profits. The US economy (a complex organization) can be thought of as generating streams of data that include the gross domestic product (GDP), stock prices, and the Consumer Price Index. Statisticians and other analysts conceptualize these data streams as being generated by processes. Typically, however, the series of operations or actions that cause particular data to be realized are either unknown or so complex (or both) that the processes are treated as *black boxes*.

A process whose operations or actions are unknown or unspecified is called a **black box.**

Frequently, when a process is treated as a black box, its inputs are not specified either. The entire focus is on the output of the process. A black box process is illustrated in Figure 1.4.

In studying a process, we generally focus on one or more characteristics, or properties, of the output. For example, we may be interested in the weight or the length of the units produced or even the time it takes to produce each unit. As with characteristics of population units, we call these characteristics *variables*. In studying processes whose output is already in numerical form (i.e., a stream of numbers), the characteristic, or property, represented by the numbers (e.g., sales, GDP, or stock prices) is typically the variable of interest. If the output is not numeric, we use *measurement processes* to assign

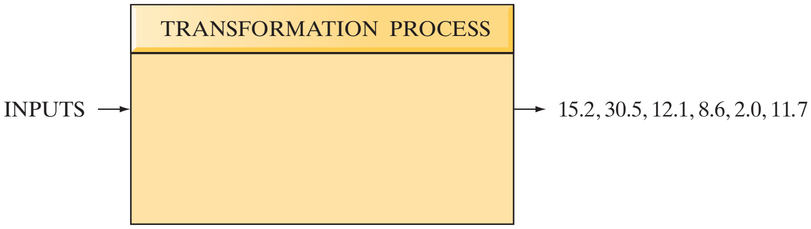


Figure 1.4
A black box process with numerical output

numerical values to variables.* For example, if in the automobile assembly process the weight of the fully assembled automobile is the variable of interest, a measurement process involving a large scale will be used to assign a numerical value to each automobile.

As with populations, we use sample data to analyze and make inferences (estimates, predictions, or other generalizations) about processes. But the concept of a sample is defined differently when dealing with processes. Recall that a population is a set of existing units and that a sample is a subset of those units. In the case of processes, however, the concept of a set of existing units is not relevant or appropriate. Processes generate or create their output *over time*—one unit after another. For example, a particular automobile assembly line produces a completed vehicle every 4 minutes. We define a sample from a process in the box.

Any set of output (object or numbers) produced by a process is also called a **sample.**

Thus, the next 10 cars turned out by the assembly line constitute a sample from the process, as do the next 100 cars or every fifth car produced today.

EXAMPLE 1.4

Key Elements of a Process—Waiting Time at a Fast-Food Window

Problem A particular fast-food restaurant chain has 6,289 outlets with drive-through windows. To attract more customers to its drive-through services, the company is considering offering a 50% discount to customers who wait more than a specified number of minutes to receive their order. To help determine what the time limit should be, the company decided to estimate the average waiting time at a particular drive-through window in Dallas, Texas. For 7 consecutive days, the worker taking customers' orders recorded the time that every order was placed. The worker who handed the order to the customer recorded the time of delivery. In both cases, workers used synchronized digital clocks that reported the time to the nearest second. At the end of the 7-day period, 2,109 orders had been timed.

a. Describe the process of interest at the Dallas restaurant.
b. Describe the variable of interest.
c. Describe the sample.
d. Describe the inference of interest.
e. Describe how the reliability of the inference could be measured.

Solution

a. The process of interest is the drive-through window at a particular fast-food restaurant in Dallas, Texas. It is a process because it "produces," or "generates," meals over time—that is, it services customers over time.

b. The variable the company monitored is customer waiting time, the length of time a customer waits to receive a meal after placing an order. Because the study is focusing only on the output of the process (the time to produce the output) and not the internal operations of the process (the tasks required to produce a meal for a customer), the process is being treated as a black box.

c. The sampling plan was to monitor every order over a particular 7-day period. The sample is the 2,109 orders that were processed during the 7-day period.

d. The company's immediate interest is in learning about the drive-through window in Dallas. They plan to do this by using the waiting times from the sample to make a statistical inference about the drive-through process. In particular, they might use the average waiting time for the sample to estimate the average waiting time at the Dallas facility.

e. As for inferences about populations, measures of reliability can be developed for inferences about processes. The reliability of the estimate of the average waiting

*A process with an output that is already in numerical form necessarily includes a measurement process as one of its subprocesses.

time for the Dallas restaurant could be measured by a bound on the error of estimation—that is, we might find that the average waiting time is 4.2 minutes, with a bound on the error of estimation of 0.5 minutes. The implication would be that we could be reasonably certain that the true average waiting time for the Dallas process is between 3.7 and 4.7 minutes.

Look Back Notice that there is also a population described in this example: the company's 6,289 existing outlets with drive-through facilities. In the final analysis, the company will use what it learns about the process in Dallas and, perhaps, similar studies at other locations to make an inference about the waiting times in its population of outlets.

Now Work Exercise 1.38

Note that output already generated by a process can be viewed as a population. Suppose a soft-drink canning process produced 2,000 twelve-packs yesterday, all of which were stored in a warehouse. If we were interested in learning something about those 2,000 twelve-packs—such as the percentage with defective cardboard packaging—we could treat the 2,000 twelve-packs as a population. We might draw a sample from the population in the warehouse, measure the variable of interest, and use the sample data to make a statistical inference about the 2,000 twelve-packs, as described in Sections 1.2 and 1.3.

In this optional section, we have presented a brief introduction to processes and the use of statistical methods to analyze and learn about processes. In Chapters 13 and 14 we present an in-depth treatment of these subjects.

1.5 Types of Data

You have learned that statistics is the science of data and that data are obtained by measuring the values of one or more variables on the units in the sample (or population). All data (and hence the variables we measure) can be classified as one of two general types: *quantitative* and *qualitative*.

Quantitative data are data that are measured on a naturally occurring numerical scale.* The following are examples of quantitative data:

1. The temperature (in degrees Celsius) at which each unit in a sample of 20 pieces of heat-resistant plastic begins to melt
2. The current unemployment rate (measured as a percentage) for each of the 50 states
3. The scores of a sample of 150 MBA applicants on the GMAT, a standardized business graduate school entrance exam administered nationwide
4. The number of female executives employed in each of a sample of 75 manufacturing companies

Quantitative data are measurements that are recorded on a naturally occurring numerical scale.

In contrast, qualitative data cannot be measured on a natural numerical scale; they can only be classified into categories.† Examples of qualitative data are as follows:

1. The political party affiliation (Democrat, Republican, or Independent) in a sample of 50 CEOs

*Quantitative data can be subclassified as either *interval* or *ratio*. For ratio data, the origin (i.e., the value 0) is a meaningful number. But the origin has no meaning with interval data. Consequently, we can add and subtract interval data, but we can't multiply and divide them. Of the four quantitative data sets listed, (1) and (3) are interval data, while (2) and (4) are ratio data.

†Qualitative data can be subclassified as either *nominal* or *ordinal*. The categories of an ordinal data set can be ranked or meaningfully ordered, but the categories of a nominal data set can't be ordered. Of the four qualitative data sets listed, (1) and (2) are nominal and (3) and (4) are ordinal.

2. The defective status (defective or not) of each of 100 computer chips manufactured by Intel
3. The size of a car (subcompact, compact, midsize, or full-size) rented by each of a sample of 30 business travelers
4. A taste tester's ranking (best, worst, etc.) of four brands of barbecue sauce for a panel of 10 testers

Often, we assign arbitrary numerical values to qualitative data for ease of computer entry and analysis. But these assigned numerical values are simply codes: They cannot be meaningfully added, subtracted, multiplied, or divided. For example, we might code Democrat = 1, Republican = 2, and Independent = 3. Similarly, a taste tester might rank the barbecue sauces from 1 (best) to 4 (worst). These are simply arbitrarily selected numerical codes for the categories and have no utility beyond that.

Qualitative data are measurements that cannot be measured on a natural numerical scale; they can only be classified into one of a group of categories.

EXAMPLE 1.5

Types of Data—Study of a River Contaminated by a Chemical Plant

Problem Chemical and manufacturing plants sometimes discharge toxic-waste materials such as DDT into nearby rivers and streams. These toxins can adversely affect the plants and animals inhabiting the river and the riverbank. The U.S. Army Corps of Engineers conducted a study of fish in the Tennessee River (in Alabama) and its three tributary creeks: Flint Creek, Limestone Creek, and Spring Creek. A total of 144 fish were captured, and the following variables were measured for each:

1. River/creek where each fish was captured
2. Species (channel catfish, largemouth bass, or smallmouth buffalo fish)
3. Length (centimeters)
4. Weight (grams)
5. DDT concentration (parts per million)

These data are saved in the **DDT** file. Classify each of the five variables measured as quantitative or qualitative.

Solution The variables length, weight, and DDT are quantitative because each is measured on a numerical scale: length in centimeters, weight in grams, and DDT in parts per million. In contrast, river/creek and species cannot be measured quantitatively: They can only be classified into categories (e.g., channel catfish, largemouth bass, and smallmouth buffalo fish for species). Consequently, data on river/creek and species are qualitative.

Look Back It is essential that you understand whether data are quantitative or qualitative in nature because the statistical method appropriate for describing, reporting, and analyzing the data depends on the data type (quantitative or qualitative).

Now Work Exercise 1.15

We demonstrate many useful methods for analyzing quantitative and qualitative data in the remaining chapters of the text. But first, we discuss some important ideas on data collection.

1.6 Collecting Data: Sampling and Related Issues

Once you decide on the type of data—quantitative or qualitative—appropriate for the problem at hand, you'll need to collect the data. Generally, you can obtain the data in three different ways:

1. Data from a *published source*
2. Data from a *designed experiment*
3. Data from an *observational study* (e.g., a *survey*)

Sometimes, the data set of interest has already been collected for you and is available in a **published source,** such as a book, journal, newspaper, or Web site. For example, you may want to examine and summarize the unemployment rates (i.e., percentages of eligible workers who are unemployed) in the 50 states of the United States. You can find this data set (as well as numerous other data sets) at your library in the *Statistical Abstract of the United States*, published annually by the US government. Similarly, someone who is interested in monthly mortgage applications for new home construction would find this data set in the *Survey of Current Business*, another government publication. Other examples of published data sources include the *Wall Street Journal* (financial data) and the The *Elias Sports Bureau* (sports information).* The Internet (World Wide Web) provides a medium by which data from published sources are readily available.

A second method of collecting data involves conducting a **designed experiment,** in which the researcher exerts strict control over the units (people, objects, or events) in the study. For example, an often-cited medical study investigated the potential of aspirin in preventing heart attacks. Volunteer physicians were divided into two groups—the *treatment* group and the *control* group. In the treatment group, each physician took one aspirin tablet a day for 1 year, while each physician in the control group took an aspirin-free placebo (no drug) made to look like an aspirin tablet. The researchers, not the physicians under study, controlled who received the aspirin (the treatment) and who received the placebo. As you will learn in Chapter 9, properly designed experiment allows you to extract more information from the data than is possible with an uncontrolled study.

Finally, observational studies can be employed to collect data. In an **observational study,** the researcher observes the experimental units in their natural setting and records the variable(s) of interest. For example, a company psychologist might observe and record the level of "Type A" behavior of a sample of assembly line workers. Similarly, a finance researcher may observe and record the closing stock prices of companies that are acquired by other firms on the day prior to the buyout and compare them to the closing prices on the day the acquisition is announced. Unlike a designed experiment, an observational study is one in which the researcher makes no attempt to control any aspect of the experimental units.

The most common type of observational study is a **survey,** where the researcher samples a group of people, asks one or more questions, and records the responses. Probably the most familiar type of survey is the political poll, conducted by any one of a number of organizations (e.g., Harris, Gallup, Roper, and CNN) and designed to predict the outcome of a political election. Another familiar survey is the Nielsen survey, which provides the major television networks with information on the most watched TV programs. Surveys can be conducted through the mail, with telephone interviews, or with in-person interviews. Although in-person interviews are more expensive than mail or telephone surveys, they may be necessary when complex information must be collected.

A **designed experiment** is a data-collection method where the researcher exerts full control over the characteristics of the experimental units sampled. These experiments typically involve a group of experimental units that are assigned the *treatment* and an untreated (or *control*) group.

An **observational study** is a data-collection method where the experimental units sampled are observed in their natural setting. No attempt is made to control the characteristics of the experimental units sampled. (Examples include *opinion polls* and *surveys*.)

*With published data, we often make a distinction between the *primary source* and *secondary source.* If the publisher is the original collector of the data, the source is primary. Otherwise, the data are secondary source.

Regardless of the data-collection method employed, it is likely that the data will be a sample from some population. And if we wish to apply inferential statistics, we must obtain a *representative sample*.

A **representative sample** exhibits characteristics typical of those possessed by the population of interest.

For example, consider a political poll conducted during a presidential election year. Assume the pollster wants to estimate the percentage of all 145 million registered voters in the United States who favor the incumbent president. The pollster would be unwise to base the estimate on survey data collected for a sample of voters from the incumbent's own state. Such an estimate would almost certainly be *biased* high; consequently, it would not be very reliable.

The most common way to satisfy the representative sample requirement is to select a simple random sample. A **simple random sample** ensures that every subset of fixed size in the population has the same chance of being included in the sample. If the pollster samples 1,500 of the 145 million voters in the population so that every subset of 1,500 voters has an equal chance of being selected, she has devised a simple random sample.

A **simple random sample** of n experimental units is a sample selected from the population in such a way that every different sample of size n has an equal chance of selection.

The procedure for selecting a simple random sample typically relies on a **random number generator.** Random number generators are available in table form, online* and in most statistical software packages. The Excel/XLSTAT, Minitab, and StatCrunch statistical software packages all have easy-to-use random number generators for creating a random sample. The next two examples illustrate the procedure.

EXAMPLE 1.6

Generating a Simple Random Sample—Selecting Households for a Feasibility Study

Problem Suppose you wish to assess the feasibility of building a new high school. As part of your study, you would like to gauge the opinions of people living close to the proposed building site. The neighborhood adjacent to the site has 711 homes. Use a random number generator to select a simple random sample of 20 households from the neighborhood to participate in the study.

Solution In this study, your population of interest consists of the 711 households in the adjacent neighborhood. To ensure that every possible sample of 20 households selected from the 711 has an equal chance of selection (i.e., to ensure a simple random sample), first assign a number from 1 to 711 to each of the households in the population. These numbers were entered into an Excel worksheet. Now, apply the random number generator of Excel/ XLSTAT, requesting that 20 households be selected without replacement. Figure 1.5 shows one possible set of random numbers generated from XLSTAT. You can see that households numbered 7, 12, 15, . . . , 704 are the households to be included in your sample.

Look Back It can be shown (proof omitted) that there are over 3×10^{38} possible samples of size 20 that can be selected from the 711 households. Random number generators guarantee (to a certain degree of approximation) that each possible sample has an equal chance of being selected.

*One of many free online random number generators is available at www.randomizer.org.

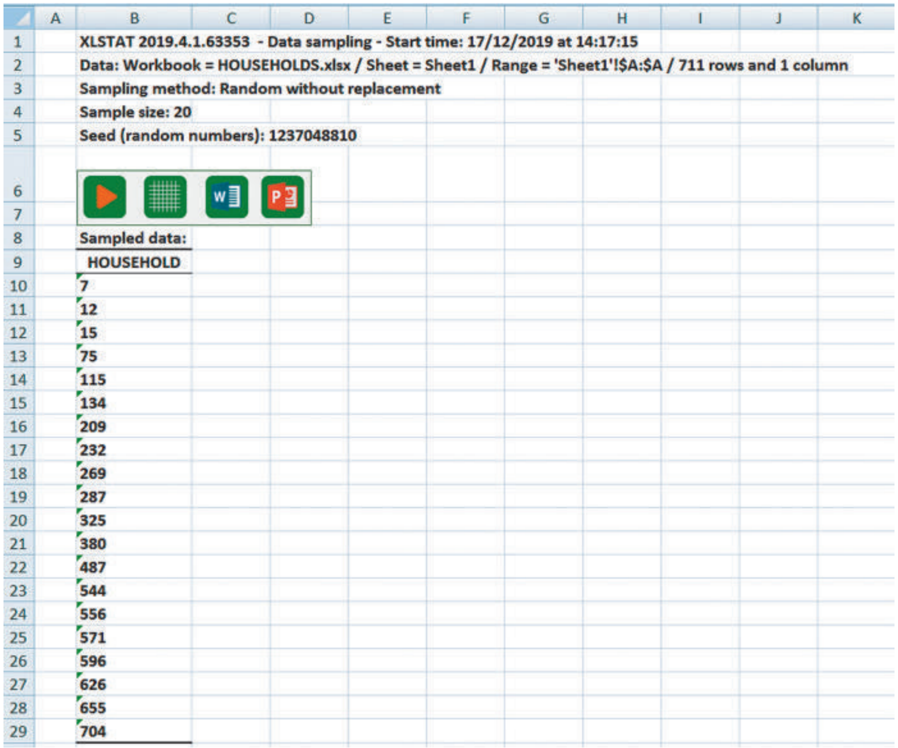


Figure 1.5
Random selection of 20 households using XLSTAT

Now Work Exercise 1.14

The notion of random selection and randomization is also key to conducting good research with a designed experiment. The next example illustrates a basic application.

EXAMPLE 1.7

Randomization in a Designed Experiment—A Clinical Trial

Problem A designed experiment in the medical field involving human subjects is referred to as a *clinical trial*. One recent clinical trial was designed to determine the potential of using aspirin in preventing heart attacks. Volunteer physicians were randomly divided into two groups—the *treatment* group and the *control* group. Each physician in the treatment group took one aspirin tablet a day for one year, while the physicians in the control group took an aspirin-free placebo made to look identical to an aspirin tablet. Because the physicians did not know which group, treatment or control, they were assigned to, the clinical trial is called a *blind study*. Assume 20 physicians volunteered for the study. Use a random number generator to randomly assign half of the physicians to the treatment group and half to the control group.

Solution Essentially, we want to select a random sample of 10 physicians from the 20. The first 10 selected will be assigned to the treatment group; the remaining 10 will be assigned to the control group. (Alternatively, we could randomly assign each physician, one by one, to either the treatment or the control group. However, this would not guarantee exactly 10 physicians in each group.)

The Minitab random sample procedure was employed, producing the printout shown in Figure 1.6. Numbering the physicians from 1 to 20, we see that physicians 1, 9, 20, 12, 3, 13, 4, 5, 14, and 15 are assigned to receive the aspirin (treatment). The remaining physicians are assigned the placebo (control).

Now Work Exercise 1.34e

↓	C1	C2
	Physician	Treatment
1	1	1
2	2	9
3	3	20
4	4	12
5	5	3
6	6	13
7	7	4
8	8	5
9	9	14
10	10	15
11	11	
12	12	
13	13	
14	14	
15	15	
16	16	
17	17	
18	18	
19	19	
20	20	
21		

Figure 1.6

Minitab Worksheet with Random Assignment of Physicians

In addition to simple random samples, there are more complex random sampling designs that can be employed. These include (but are not limited to) **stratified random sampling**, **cluster sampling**, **systematic sampling,** and **randomized response sampling.** Brief descriptions of each follow. (For more details on the use of these sampling methods, consult the references at the end of this chapter.)

Stratified random sampling is typically used when the experimental units associated with the population can be separated into two or more groups of units, called *strata*, where the characteristics of the experimental units are more similar within strata than across strata. Random samples of experimental units are obtained for each strata; then the units are combined to form the complete sample. For example, if you are gauging opinions of voters on a polarizing issue, like government-sponsored health care, you may want to stratify on political affiliation (Republicans and Democrats), making sure that representative samples of both Republicans and Democrats (in proportion to the number of Republicans and Democrats in the voting population) are included in your survey.

Sometimes it is more convenient and logical to sample natural groupings (*clusters*) of experimental units first, and then collect data from all experimental units within each cluster. This involves the use of *cluster sampling*. For example, suppose a marketer for a large upscale restaurant chain wants to find out whether customers like the new menu. Rather than collect a simple random sample of all customers (which would be very difficult and costly to do), the marketer will randomly sample 10 of the 150 restaurant locations (clusters), and then interview all customers eating at each of the 10 locations on a certain night.

Another popular sampling method is *systematic sampling*. This method involves systematically selecting every *k*th experimental unit from a list of all experimental units. For example, every fifth person who walks into a shopping mall could be asked his or her opinion on a business topic of interest. Or, a quality control engineer at a manufacturing plant may select every 10th item produced on an assembly line for inspection.

A fourth alternative to simple random sampling is *randomized response sampling*. This design is particularly useful when the questions of the pollsters are likely to elicit false answers. For example, suppose each person in a sample of wage earners is asked whether he or she ever cheated on an income tax return. A cheater might lie, thus biasing an estimate of the true likelihood of someone cheating on his or her tax return. To circumvent this problem, each person is presented with two questions, one being the object of the survey and the other an innocuous question, such as:

1. Did you ever cheat on your federal income tax return?
2. Did you drink coffee this morning?

One of the questions is chosen at random to be answered by the wage earner by flipping a coin; however, which particular question is answered is unknown to the interviewer. In this way, the random response method attempts to elicit an honest response to a sensitive question. Sophisticated statistical methods are then employed to derive an estimate of the percentage of "yes" responses to the sensitive question.

No matter what type of sampling design you employ to collect the data for your study, be careful to avoid *selection bias*. Selection bias occurs when some experimental units in the population have less chance of being included in the sample than others. This results in samples that are not representative of the population. Consider an opinion poll that employs either a telephone survey or a mail survey. After collecting a random sample of phone numbers or mailing addresses, each person in the sample is contacted via telephone or the mail and a survey conducted. Unfortunately, these types of surveys often suffer from selection bias due to *nonresponse*. Some individuals may not be home when the phone rings, or others may refuse to answer the questions or mail back the questionnaire. As a consequence, no data are obtained for the nonrespondents in the sample. If the nonrespondents and respondents differ greatly on an issue, then *nonresponse bias* exits. For example, those who choose to answer a question on a school board issue may have a vested interest in the outcome of the survey—say, parents with children of school age, schoolteachers whose jobs may be in jeopardy, or citizens whose taxes might be substantially affected. Others with no vested interest may have an opinion on the issue but might not take the time to respond.

Selection bias results when a subset of experimental units in the population has little or no chance of being selected for the sample.

Consider a sample of experimental units where some units produce data (i.e., responders) and no data is collected on the other units (i.e., nonresponders). **Nonresponse bias** is a type of selection bias that results when the response data differ from the potential data for the nonresponders.

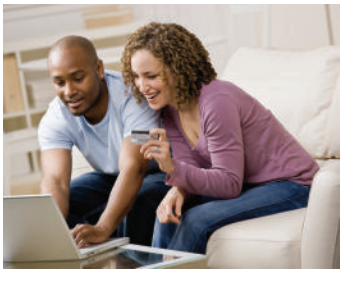
Finally, even if your sample is representative of the population, the data collected may suffer from *measurement error*. That is, the values of the data (quantitative or qualitative) may be inaccurate. In sample surveys, opinion polls, etc., measurement error often results from *ambiguous* or *leading questions*. Consider the survey question: "How often did you change the oil in your car last year?" It is not clear whether the researcher wants to know how often you personally changed the oil in your car or how often you took your car into a service station to get an oil change. The ambiguous question may lead to inaccurate responses. On the other hand, consider the question: "Does the new health plan offer more comprehensive medical services at less cost than the old one?" The way the question is phrased *leads* the reader to believe that the new plan is better and to a "yes" response—a response that is more desirable to the researcher. A better, more neutral way to phrase the question is: "Which health plan offers more comprehensive medical services at less cost, the old one or the new one?"

Measurement error refers to inaccuracies in the values of the data collected. In surveys, the error may be due to ambiguous or leading questions and the interviewer's effect on the respondent.

We conclude this section with two examples involving actual sampling studies.

EXAMPLE 1.8

Method of Data Collection—Survey of Online Shoppers



Problem What is the most popular device used by online shoppers? To find out, the mobile video ad network AdColony conducted a 2019 nationwide survey of 1,000 US online shoppers for Mobile Marketer. The most popular device was a smartphone, used by 56% of the online shoppers. Other results: 28% used a desktop or laptop computer, and 16% used a tablet.

a. Identify the data-collection method.
b. Identify the target population.
c. Are the sample data representative of the population?

Solution

a. The data-collection method is a survey: 1,000 online shoppers participated in the study.

b. Presumably, Mobile Marketer (who commissioned the survey) is interested in the devices used by all US online shoppers. Consequently, the target population is *all* consumers who use the Internet for online shopping.

c. Because the 1,000 respondents clearly make up a subset of the target population, they do form a sample. Whether or not the sample is representative is unclear because Mobile Marketer provided no detailed information on how the 1,000 shoppers were selected. If the respondents were obtained using, say, random-digit telephone dialing, then the sample is likely to be representative because it is a random sample. However, if the questionnaire was made available to anyone surfing the Internet, then the respondents are *self-selected* (i.e., each Internet user who saw the survey chose whether or not to respond to it). Such a survey often suffers from *nonresponse* bias. It is possible that many Internet users who chose not to respond (or who never saw the questionnaire) would have answered the questions differently, leading to a lower (or higher) sample percentage.

Look Back Any inferences based on survey samples that employ self-selection are suspect due to potential nonresponse bias.

Now Work Exercise 1.27

EXAMPLE 1.9

Representative Data—Price Promotion Study

Problem Marketers use wording such as "was \$100, now \$80" to indicate a price promotion. The promotion is typically compared to the retailer's previous price or to a competitor's price. A study in the *Journal of Consumer Research* investigated whether between-store comparisons result in greater perceptions of value by consumers than within-store comparisons. Suppose 50 consumers were randomly selected from all consumers in a designated market area to participate in the study. The researchers randomly assigned 25 consumers to read a within-store price promotion advertisement ("was \$100, now \$80") and 25 consumers to read a between-store price promotion ("\$100 there, \$80 here"). The consumers then gave their opinion on the value of the discount offer on a 10-point scale (where $1 =$ lowest value and $10 =$ highest value). The value opinions of the two groups of consumers were compared.

a. Identify the data-collection method.

b. Are the sample data representative of the target population?

Solution

a. Here, the experimental units are the consumers. Because the researchers controlled which price promotion ad—"within-store" or "between-store"—the experimental units (consumers) were assigned to, a designed experiment was used to collect the data.

b. The sample of 50 consumers was randomly selected from all consumers in the designated market area. If the target population is all consumers in this market, it is likely that the sample is representative. However, the researchers warn that the sample data should not be used to make inferences about consumer behavior in other, dissimilar markets.

Look Back By using randomization in a designed experiment, the researcher is attempting to eliminate different types of bias, including self-selection bias.

Now Work Exercise 1.19

STATISTICS IN ACTION REVISITED

Identifying the Data-Collection Method and Data Type

First, refer to the *Nature Neuroscience* chocolate study. Recall that researchers randomly assigned 37 people to one of four groups: (1) High dose of cocoa flavanol and aerobic exercise; (2) High dose of cocoa flavanol but no exercise; (3) Low dose of cocoa flavanol and aerobic exercise; and (4) Low cocoa flavanol dose but without the exercise (See Figure SIA1.1). After a 3-month period, the researchers measured (among other variables) each subject's reaction time (in seconds) to an object-recognition task. Clearly, the data collection method employed is a designed experiment—the experimental units (subjects) were randomly assigned to the groups. The groups are formed from two variables: dosage of cocoa flavanol (high or low) and aerobic exercise (yes or no). Consequently, these two categorical-type variables are qualitative in nature. The numerical reaction time variable, measured in seconds, is quantitative.

Now, let's consider the *New York Times* study on the link between a CEO's golf handicap and the company's stock performance. Recall that the newspaper gathered information on golf handicaps of corporate executives obtained from a *Golf Digest* survey

STATISTICS IN ACTION REVISTED (*continued*)

that was sent to 300 corporate executives. Thus, the data-collection method is a survey. In addition to golf handicap (a numerical "index" that allows golfers to compare skills), the *Times* measured the CEO's company stock market performance over a 3-year period on a scale of 0 to 100. Because both variables, golf handicap and stock performance, are numerical in nature, they are quantitative data.

Data Set: GLFCEO

1.7 Business Analytics: Critical Thinking with Statistics

BIOGRAPHY

H. G. WELLS (1866–1946)
Writer and Novelist

English-born Herbert George Wells published his first novel, *The Time Machine*, in 1895 as a parody of the English class division and as a satirical warning that human progress is inevitable. Although most famous as a science-fiction novelist, Wells was a prolific writer as a journalist, sociologist, historian, and philosopher. Wells's prediction about statistical thinking is just one of a plethora of observations he made about life in this world. Here are a few more of H. G. Wells's more famous quotes:

"Advertising is legalized lying."

"Crude classification and false generalizations are the curse of organized life."

"The crisis of today is the joke of tomorrow."

"Fools make researchers and wise men exploit them."

"The only true measure of success is the ratio between what we might have done and what we might have been on the one hand, and the thing we have made and the things we have made of ourselves on the other." (Quotes by Herbert George Wells.)

According to H. G. Wells, author of such science-fiction classics as *The War of the Worlds* and *The Time Machine, "Statistical thinking* will one day be as necessary for efficient citizenship as the ability to read and write." Written more than a hundred years ago, Wells's prediction is proving true today.

The growth in data collection associated with scientific phenomena, business operations, and government activities (e.g., marketing, quality control, statistical auditing, forecasting, etc.) has been remarkable over the past decade. This growth is due, in part, to technology now capable of capturing lots of data at a high rate, such as information-sensing mobile devices, cameras, radio-frequency identification (RFID) readers, and wireless sensor networks. In fact, the term "Big Data" is now commonly used by companies to describe this wealth of information.

However, with big data, comes the need for methods of analysis—**business analytics**—that ultimately lead to good business decisions. A key to successful applications of business analytics is *quantitative literacy* (i.e., the ability to evaluate data intelligently). Whether the data of interest is "big" or not, each of us has to develop the ability to use rational thought to interpret and understand the meaning of the data. Business analytics and quantitative literacy can help you make intelligent decisions, inferences, and generalizations from data; that is, it helps you *think critically* using statistics. We term this skill **statistical thinking.**

Business analytics refers to methodologies (e.g., statistical methods) that extract useful information from data in order to make better business decisions.

Statistical thinking involves applying rational thought and the science of statistics to critically assess data and inferences. Fundamental to the thought process is that variation exists in populations and process data.

To gain some insight into the role statistics plays in business analytics, we present two examples of some misleading or faulty surveys.

EXAMPLE 1.10

Biased Sample—Motorcycle Helmet Law

Problem An article in the *New York Times* considered the question of whether motorcyclists should be required by law to wear helmets. In supporting his argument for no helmets, the editor of a magazine for Harley-Davidson bikers presented the results of one study that claimed "nine states without helmet laws had a lower fatality rate (3.05 deaths per 10,000 motorcycles) than those that mandated helmets (3.38)" and a survey that found "of 2,500 bikers at a rally, 98% of the respondents opposed such laws." Based on this information, do you think it is safer to ride a motorcycle without a helmet? What further statistical information would you like?

Solution You can use statistical thinking to help you critically evaluate the study. For example, before you can evaluate the validity of the 98% estimate, you would want to know how the data were collected. If a survey was, in fact, conducted, it's possible that the 2,500 bikers in the sample were not selected at random from the target population of all bikers, but rather were "self-selected." (Remember, they were all attending a rally—a rally likely for bikers who oppose the law.) If the respondents were likely to have strong opinions regarding the helmet law (e.g., strongly oppose the law), the resulting estimate is probably biased high. Also, if the selection bias in the sample was intentional, with the sole purpose to mislead the public, the researchers would be guilty of **unethical statistical practice.**

ETHICS in STATISTICS

Intentionally selecting a biased sample in order to produce misleading statistics is considered unethical statistical practice.

You would also want more information about the study comparing the motorcycle fatality rate of the nine states without a helmet law to those states that mandate helmets. Were the data obtained from a published source? Were all 50 states included in the study, or were only certain states selected? That is, are you seeing sample data or population data? Furthermore, do the helmet laws vary among states? If so, can you really compare the fatality rates?

Look Back Questions such as these led a group of mathematics and statistics teachers attending an American Statistical Association course to discover a scientific and statistically sound study on helmets. The study reported a dramatic *decline* in motorcycle crash deaths after California passed its helmet law.

EXAMPLE 1.11

Manipulative or Ambiguous Survey Questions—Satellite Radio Survey

Problem When talk-show host Howard Stern moved his controversial radio program from free, over-the-air (AM/FM) radio to Sirius XM satellite radio, the move was perceived in the industry to boost satellite radio subscriptions. This led American Media Services, a developer of AM/FM radio properties, to solicit a nationwide random-digit dialing phone survey of 1,008 people. The purpose of the survey was to determine how much interest Americans really have in buying satellite radio service. After providing some background on Howard Stern's controversial radio program, one of the questions asked, "How likely are you to purchase a subscription to satellite radio after Howard Stern's move to Sirius?" The result: 86% of the respondents stated that they aren't likely to buy satellite radio because of Stern's move. Consequently, American Media Services concluded that "the Howard Stern Factor is overrated" and that "few Americans expect to purchase satellite radio"—claims that made the headlines of news reports and Web blogs. Do you agree?

Solution First, we need to recognize that American Media Services had a vested interest in the outcome of the survey—the company makes its money from over-the-air broadcast radio stations. Second, although the phone survey was conducted using random-digit dialing, there is no information provided on the response rate. It's possible that nonrespondents (people who were not home or refused to answer the survey questions) tend to be people who use cell phones more than their landline phone and, consequently, are more likely to use the latest in electronic technology, including satellite radio. Finally, the survey question itself is ambiguous. Do the respondents have negative feelings about satellite radio, Howard Stern, or both? If not for Howard Stern's program, would the respondents be more likely to buy satellite radio? To the critical thinker, it's unclear what the results of the survey imply.

Look Back Examining the survey results from the perspective of satellite radio providers, 14% of the respondents indicated that they would be likely to purchase satellite radio. Projecting the 14% back to the population of all American adults, this figure represents about 50 million people; what is interpreted as "few Americans" by American Media Services could be music to the ears of satellite radio providers.

ETHICS in STATISTICS

Intentionally selecting a nonrandom sample in an effort to support a particular viewpoint is considered *unethical statistical practice.*

As with many statistical studies, both the motorcycle helmet study and the satellite radio study are based on survey data. Most of the problems with these surveys result from the use of *nonrandom* samples. These samples are subject to potential errors, such as *selection bias, nonresponse bias*, and *measurement error*. Researchers who are aware of these problems and continue to use the sample data to make inferences are practicing *unethical statistics*.

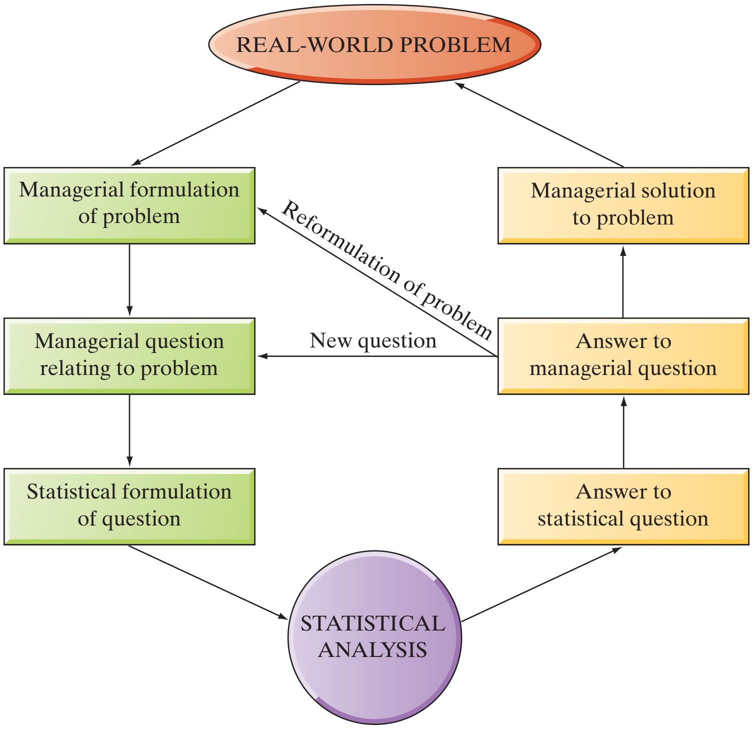


Figure 1.7
Flow diagram showing the role of statistics in business analytics
Source: From *The American Statistician* by George Benson. Copyright © by George Benson. Used by permission of George Benson.

As stated earlier, business analytics relies heavily on statistical thinking to help firms make better business decisions. The role statistics can play in a manager's use of business analytics is displayed in Figure 1.7. Every managerial decision-making problem begins with a real-world problem. This problem is then formulated in managerial terms and framed as a managerial question. The next sequence of steps (proceeding counterclockwise around the flow diagram) identifies the role that statistics can play in this process. The managerial question is translated into a statistical question, the sample data are collected and analyzed, and the statistical question is answered. The next step in the process is using the answer to the statistical question to reach an answer to the managerial question. The answer to the managerial question may suggest a reformulation of the original managerial problem, suggest a new managerial question, or lead to the solution of the managerial problem.

One of the most difficult steps in the decision-making process—one that requires a cooperative effort among managers and statisticians—is the translation of the managerial question into statistical terms (for example, into a question about a population). This statistical question must be formulated so that, when answered, it will provide the key to the answer to the managerial question. Thus, as in the game of chess, you must formulate the statistical question with the end result, the solution to the managerial question, in mind.

In the remaining chapters of the text, you'll become familiar with the business analytic tools essential for building a firm foundation in statistics and statistical thinking.

STATISTICS IN ACTION REVISITED

Critically Assessing the Ethics of a Statistical Study

The *New York Times* reported a strong link between a corporate executive's golf handicap and his/her company's stock performance. Thus, the newspaper inferred that the better the CEO is at golf, the better the company's stock performance will be. To critically assess this study, consider the following facts:

1. *Golf Digest* sent surveys to the CEOs at the 300 largest US firms. Only 74 executives agreed to reveal their golf handicaps. Of these 74 CEOs, the *Times* collected data on stock performance for only 51 of the companies. (The other 23 CEOs were not in the stock performance database used by the newspaper.)

STATISTICS IN ACTION REVISITED *(continued)*

2. The *New York Times* researcher who performed the analysis of the data stated that "for all the different factors I've tested as possible links to predicting which CEOs are going to perform well or poorly, [golf handicap] is certainly one of the . . . strongest."
3. According to the *Times*, the researcher "scientifically sifted out a handful of CEOs because of their statistical extremes," in effect "removing seven CEOs from the final analysis because [their data] destroyed the trend lines."

These observations lead a critical thinker to doubt the validity of the inference made by the *New York Times* researcher. Consider first that the sample of CEOs analyzed was not randomly selected from all CEOs in the United States. In fact, it was self-selected—only those CEOs who chose to report their golf handicap were included in the study. (Not even all these "self-reporters" were included; some were eliminated because the newspaper did not have information on their company's stock performance in the database.) Thus, the potential for selection and/or nonresponse bias is extremely high.

Second, based on fact #2, it is likely that the researcher tested a multitude of factors and found only one (golf handicap) that had a link to stock performance. We will learn in subsequent chapters that a plethora of irrelevant variables are tested statistically, by chance one or more of the variables will be found "statistically significant."

Finally, the researcher removed the data for seven CEOs based on their "statistical extremes." In the next chapter, we learn about statistical "outliers"—how to detect them and how to treat them when discovered. However, it can be shown (using the methods outlined in the text) that these seven data points are not outliers. If the data points are included in the analysis, the link between golf handicap and stock performance is found to be weak, at best.

Data Set: GLFCEO

CHAPTER NOTES

Key Terms

Note: Starred (*) terms are from the optional section in this chapter.

*Black box 12
Big data 22
Business analytics 22
Census 7
Cluster sampling 19
Data 3
Descriptive statistics 5
Designed experiment 16
Experimental (or observational) unit 7
Inference 5
Inferential statistics 5
Measurement 7
Measurement error 19
Measure of reliability 10
Nonresponse bias 20
Observational study 16
Population 7
*Process 11
Published source 16
Qualitative data 15
Quantitative data 14
Quantitative literacy 21
Randomized response sampling 19
Random number generator 17
Reliability 10
Representative sample 17
Sample 8
Selection bias 20
Simple random sample 17
Statistical inference 8
Statistical thinking 22
Statistics 4
Stratified random sampling 19
Survey 16
Systematic sampling 19
Unethical statistical practice 23
Variable 7

Key Ideas

Types of Statistical Applications

Descriptive

1. Identify **population** or **sample** (collection of **experimental units**)
2. Identify **variable(s)**
3. Collect **data**
4. **Describe** data

Inferential

1. Identify **population** (collection of *all* **experimental units**)
2. Identify **variable(s)**
3. Collect **sample** data (*subset* of population)
4. **Inference** about population based on sample
5. **Measure of reliability** for inference

Types of Data

1. **Quantitative** (numerical in nature)
2. **Qualitative** (categorical in nature)

Data-Collection Methods

1. **Observational** (e.g., survey)
2. **Published source**
3. **Designed experiment**

Types of Random Samples

1. **Simple random sample**
2. **Stratified random sample**
3. **Cluster sample**
4. **Systematic sample**
5. **Random response sample**

Problems with Nonrandom Samples

1. **Selection bias**
2. **Nonresponse bias**
3. **Measurement error**

EXERCISES 1.1–1.40

Note: Starred (*) exercises are from the optional section in this chapter.

Learning the Mechanics

1.1 What is statistics?

1.2 Explain the difference between descriptive and inferential statistics.

1.3 List and define the four elements of a descriptive statistics problem.

1.4 List and define the five elements of an inferential statistical analysis.

1.5 List the three major methods of collecting data and explain their differences.

1.6 Explain the difference between quantitative and qualitative data.

1.7 Explain how populations and variables differ.

1.8 Explain how populations and samples differ.

1.9 What is a representative sample? What is its value?

1.10 Why would a statistician consider an inference incomplete without an accompanying measure of its reliability?

***1.11** Explain the difference between a population and a process.

1.12 Define *statistical thinking*.

1.13 Suppose you're given a data set that classifies each sample unit into one of four categories: A, B, C, or D. You plan to create a computer database consisting of these data, and you decide to code the data as A = 1, B = 2, C = 3, and D = 4. Are the data consisting of the classifications A, B, C, and D qualitative or quantitative? After the data are input as 1, 2, 3, or 4, are they qualitative or quantitative? Explain your answers.

1.14 NW Suppose that a population contains 200,000 experimental units. Use a random number generator to select a simple random sample of $n = 10$ units from the population.

Applet Exercise 1.1

The *Random Numbers* applet generates a list of n random numbers from 1 to N, where n is the size of the sample and N is the size of the population. The list generated often contains repetitions of one or more numbers.

a. Using the applet *Random Numbers*, enter 1 for the minimum value, 10 for the maximum value, and 10 for the sample size. Then click on *Sample*. Look at the results and list any numbers that are repeated and the number of times each of these numbers occurs.

b. Repeat part **a,** changing the maximum value to 20 and keeping the size of the sample fixed at 10. If you still have repetitions, repeat the process, increasing the maximum value by 10 each time but keeping the size of the sample fixed. What is the smallest maximum value for which you had no repetitions?

c. Describe the relationship between the population size (maximum value) and the number of repetitions in the list of random numbers as the population size increases and the sample size remains the same. What can you conclude about using a random number generator to choose a relatively small sample from a large population?

Applet Exercise 1.2

The *Random Numbers* applet can be used to select a random sample from a population, but can it be used to simulate data? In parts **a** and **b**, you will use the applet to create data sets. Then you will explore whether those data sets are realistic.

a. In the activity *Keep the Change* on page 30, a data set called *Amounts Transferred* is described. Use the *Random Numbers* applet to simulate this data set by setting the minimum value equal to 0, the maximum value equal to 99, and the sample size equal to 30. Explain what the numbers in the list produced by the applet represent in the context of the activity. (You may need to read the activity.) Do the numbers produced by the applet seem reasonable? Explain.

b. Use the *Random Numbers* applet to simulate grades on a statistics test by setting the minimum value equal to 0, the maximum value equal to 100, and the sample size equal to 30. Explain what the numbers in the list produced by the applet represent in this context. Do the numbers produced by the applet seem reasonable? Explain.

c. Referring to parts **a** and **b,** why do the randomly generated data seem more reasonable in one situation than in the other? Comment on the usefulness of using a random number generator to produce data.

Applying the Concepts—Basic

1.15 NW **Wind turbine database**. A wind turbine turns wind energy into electricity using the aerodynamic force from the rotor blades. In 2019, wind power accounted for 5 percent of all electricity generated in the United States. The US Geological Survey compiled a database of over 53,700 wind turbines in the country. Several variables measured for each wind turbine in the database are listed below. Determine the type of data (quantitative or qualitative) recorded for each variable.

a. Electrical generation capacity (measured in kilowatts)
b. Hub height (measured in meters)
c. Rotor diameter (measured in meters)
d. Location (state/county)
e. Number of turbines in the project

[*Note*: A project consists of multiple wind turbines that form a single system.]

1.16 NW CAPRATE **Valuation of single-tenant properties**. In court cases involving real estate, experts are hired to estimate the value of properties. The estimate will vary depending on the method of valuation applied. Research on valuation methods for single-tenant properties was published in *The Appraisal Journal* (Summer 2019). One method examines the ratio of net operating income to property asset value – called the capitalization rate. The research focused on 80 tenants of properties owned by the Boulder Group consulting firm. Data on S&P credit ratings and capitalization rates for the 13 Boulder Group tenants of retail properties are shown in the table (next page). Capitalization rates are provided for tenants with 5, 10, 15, and 20 years remaining on the lease.

a. What is the experimental unit for this study?
b. Describe the variables measured in the study. Do these variables produce quantitative or qualitative data?

c. Do the 13 tenants in the table represent a population or a sample?
d. If the 13 tenants represent a representative sample from a population, describe the population.
e. One inference derived from the data is that "capitalization rates increase . . . as the remaining lease term decreases." Do you agree?

Tenant	S&P Credit Rating	Capitalization Rate (%) 5-Year	10-Year	15-Year	20-Year
Best Buy	BBB	8.25	7.25	N/A	N/A
BJ's Warehouse	B	7.40	6.85	6.40	6.00
Dollar General	BBB−	8.45	7.75	6.90	N/A
Dollar Tree	BBB−	7.50	6.85	N/A	N/A
Family Dollar	BBB−	8.50	7.50	N/A	N/A
Kohl's	BBB−	8.50	7.25	6.50	6.35
Kroger	BBB	7.25	6.75	6.25	6.00
Lowe's	BBB+	6.70	6.10	5.45	5.00
Sherwin-Williams	BBB	6.75	5.90	N/A	N/A
The Home Depot	A	6.75	6.00	5.40	5.00
United Rentals	BB	8.15	7.25	N/A	N/A
Walmart	AA	6.75	5.90	5.50	5.15
Whole Foods Market	AA−	5.50	5.15	4.60	4.25

Source: Sellers, L.P., et al. "Valuation Methods and Dark Big-Box Theories", *The Appraisal Journal*, Vol. 87, No. 3, Summer 2019 (Exhibit 1).

1.17 Disaggregation of annually reported accounting data. Researchers have constructed a new measure of disclosure quality—*called disaggregation quality (DQ)*—based on the level of disaggregation of financial data items in a firm's annual reports (*Journal of Accounting Research,* December 2015). Quantitative DQ values were computed for each of 125,873 firms that had annual balance sheet data from 1973 to 2011 reported on Compustat. One objective is to describe the distribution of the DQ values for firms reported on Compustat.
a. What is the experimental unit for this study?
b. What is the variable measured in this study?
c. Are the data collected from a designed experiment or an observational study? Explain.

1.18 Jamming attacks on wireless networks. Terrorists often use wireless networks to communicate. To disrupt these communications, the US military uses jamming attacks on the wireless networks. The *International Journal of Production Economics* (Vol. 172, 2016) described a study of 80 such jamming attacks. The configuration of the wireless network attacked was determined in each case. Configuration consists of network type (WLAN, WSN, or AHN) and number of channels (single- or multi-channel).
a. Suppose the 80 jamming attacks represent all jamming attacks by the US military over the past several years, and these attacks are the only attacks of interest to the researchers. Do the data associated with these 80 attacks represent a population or a sample? Explain.
b. The 80 jamming attacks actually represent a sample. Describe the population for which this sample is representative.
c. Identify the variable "network type" as quantitative or qualitative.
d. Identify the variable "number of channels" as quantitative or qualitative.
e. Explain how to measure number of channels quantitatively.

1.19 NW **Opinion polls.** Pollsters regularly conduct opinion polls to determine the popularity rating of the current president. Suppose a poll is to be conducted tomorrow in which 2,000 individuals will be asked whether the president is doing a good or bad job. The 2,000 individuals will be selected by random-digit telephone dialing and asked the question over the phone.
a. What is the relevant population?
b. What is the variable of interest? Is it quantitative or qualitative?
c. What is the sample?
d. What is the inference of interest to the pollster?
e. What method of data collection is employed?
f. How likely is the sample to be representative?

1.20 Cybersecurity survey. The information systems organization, ISACA, conducts an annual survey of cybersecurity at firms from around the world. ISACA sends survey questionnaires via email (SurveyMonkey) to all professionals that hold ISACA's Certified Information Security Manager designation. Over 1,500 professionals participated in the 2019 cybersecurity survey. Each was asked whether or not they expect to experience a cyberattack (e.g., a Malware, hacking, or phishing attack) against their firm in the coming year. About 80% of the respondents expect to experience a cyberattack during the year (*State of Cybersecurity: 2019: Part 2*, ISACA and RSA Conference Survey).
a. Identify the population of interest to ISACA.
b. Identify the data-collection method used by ISACA. Are there any potential biases in the method used?
c. Describe the variable measured in the ISACA survey. Is it quantitative or qualitative?
d. What inference can be made from the study result?

1.21 Treasury deficit prior to the Civil War. In *Civil War History* (June 2009), historian Jane Flaherty researched the condition of the US Treasury on the eve of the Civil War in 1861. Between 1854 and 1857 (under President Franklin Pierce), the annual surplus/deficit was +18.8, +6.7, +5.3, and +1.3 million dollars, respectively. In contrast, between 1858 and 1861 (under President James Buchanan), the annual surplus/deficit was −27.3, −16.2, −7.2, and −25.2 million dollars, respectively. Flaherty used these data to aid in portraying the exhausted condition of the U.S. Treasury when Abraham Lincoln took office in 1861. Does this study represent a descriptive or inferential statistical study? Explain.

1.22 The "lucky store effect" in lottery ticket sales. In the *American Economic Review* (Vol. 98, 2008), University of Chicago researchers investigated the *lucky store effect* theory in lottery ticket sales, i.e., the theory that a lottery retail store that sold a large-prize-winning ticket will experience greater ticket sales the following week. The researchers examined the weekly ticket sales of all 24,400 active lottery retailers in Texas. The analysis showed that "the week following the sale of [a winning Lotto Texas ticket], the winning store experiences a 12 to 38 percent relative sales increase. . . . " Consequently, the researchers project that future winning lottery retail stores will experience the *lucky store effect*. Is this study an example of descriptive statistics or inferential statistics? Explain.

Applying the Concepts—Intermediate

1.23 Delivery times for online orders. The *Journal of Marketing Research* (October 2019) published a study of delivery times for online orders. During a recent year, a major apparel retailer fulfilled all its online orders from a single distribution center (DC) located in the eastern United States. Later that year, the retailer opened a second DC located in the western United States, with the goal of reducing delivery times to western US customers. The researchers collected data on delivery times for a sample of online orders fulfilled by both the eastern and western US distribution centers for several western states. For the state of Washington, the typical delivery time was 7 business days from the eastern DC and 5 business days from the western DC. For the state of Montana, the typical delivery time was 7 business days from either of the distribution centers.

a. What is the experimental unit for this study?
b. Identify the variables measured and their type (quantitative or qualitative).
c. Explain why this is an example of inferential statistics. What inference can you make?

1.24 Who is better at multi-tasking? In business, employees are often asked to perform a complex task when their attention is divided (i.e., *multi-tasking). Human Factors* (May 2014) published a study designed to determine whether video game players are better than non–video game players at multi-tasking. Each in a sample of 60 college students was classified as a video game player or a non-player. Participants entered a street crossing simulator and were asked to cross a busy street at an unsigned intersection. The simulator was designed to have cars traveling at various high rates of speed in both directions. During the crossing, the students also performed a memory task as a distraction. Two variables were measured for each student: (1) a street crossing performance score (measured out of 100 points) and (2) a memory task score (measured out of 20 points). The researchers found no differences in either the street crossing performance or memory task score of video game players and non-gamers. "These results," say the researchers, "suggest that action video game players [and non-gamers] are equally susceptible to the costs of dividing attention in a complex task."

a. Identify the experimental unit for this study.
b. Identify the variables measured as quantitative or qualitative.
c. Is this an application of descriptive statistics or inferential statistics? Explain.

1.25 Consumer recycling behavior. Under what conditions will consumers dispose of recyclable paper in the garbage? This was the question of interest in an article published in the *Journal of Consumer Research* (December, 2013). In one of the studies conducted, the researchers instructed 78 college students to cut an 8.5-×11-inch sheet of paper into eight smaller pieces. Half the students were randomly assigned to list five uses for the cut paper (*usefulness is salient* condition). The students in the other half were asked to list their five favorite TV shows (*control condition*). After completing an unrelated task, all students were asked to dispose of the paper upon leaving. There was a trash bin and a recycling bin outside the door. The researchers kept track of which students recycled and which students disposed of their paper in the garbage. This information was used to test the theory that students in the *usefulness is salient* condition will recycle at a higher rate than students in the *control* condition.

a. Explain why the data-collection method used in this study is a designed experiment.
b. Identify the experimental unit in this study.
c. Identify the variables measured in this study. What type of data (quantitative or qualitative) is produced from each variable? (*Hint*: Two variables are measured.)
d. About 68% of the students in the *usefulness is salient* condition recycled, as compared to 37% of students in the *control* condition. Use this information to make an inference about the population of interest.

1.26 Drafting NFL quarterbacks. The National Football League (NFL) is a lucrative business, generating an annual revenue of about $15 billion. One key to becoming a financially successful NFL team is drafting a good quarterback (QB) out of college. The NFL draft allows the worst-performing teams in the previous year the opportunity of selecting the best quarterbacks coming out of college. The *Journal of Productivity Analysis* (Vol. 35, 2011) published a study of how successful NFL teams are in drafting productive quarterbacks. Data were collected for 331 quarterbacks drafted over a 38-year period. Several variables were measured for each QB, including draft position (one of the top 10 players picked, selection between picks 11 and 50, or selected after pick 50), NFL winning ratio (percentage of games won), and QB production score (higher scores indicate more productive QBs). The researchers discovered that draft position is only weakly related to a quarterback's performance in the NFL. They concluded that "quarterbacks taken higher [in the draft] do not appear to perform any better."

a. What is the experimental unit for this study?
b. Identify the type (quantitative or qualitative) of each variable measured.
c. Suppose you want to use this study to project the performance of future NFL QBs. Is this an application of descriptive or inferential statistics? Explain.

1.27 NW **The economic return to earning an MBA.** What are the economic rewards (e.g., higher salary) to obtaining an MBA degree? This was the question of interest in an article published in the *International Economic Review* (August 2008). The researchers made inferences based on wage data collected for a sample of 3,244 individuals who sat for the Graduate Management Admissions Test (GMAT). (The GMAT exam is required for entrance into most MBA programs.) The following sampling scheme was employed. All those who took the GMAT exam in any of four selected time periods were mailed a questionnaire. Those who responded to the questionnaire were then sent three follow-up surveys (one survey every 3 months). The final sample of 3,244 represents only those individuals who responded to all four surveys. (For example, about 5,600 took the GMAT in one time period; of these, only about 800 responded to all four surveys.)

a. For this study, describe the population of interest.
b. What method was used to collect the sample data?
c. Do you think the final sample is representative of the population? Why or why not? Comment on potential biases in the sample.

1.28 Corporate sustainability and firm characteristics. *Corporate sustainability* refers to business practices designed around social and environmental considerations (e.g., "going green"). *Business and Society* (March 2011) published a paper on how firm size and firm type impact sustainability behaviors. The researchers added questions on sustainability to a quarterly survey of Certified Public Accountants (CPAs). The survey was sent to approximately 23,500 senior managers at CPA firms, of which 1,293 senior managers responded. (*Note:* It is not clear how the 23,500 senior managers were selected.) Due to missing data (incomplete survey answers), only 992 surveys were analyzed. These data were used to infer whether larger firms are more likely to report sustainability policies than smaller firms and whether public firms are more likely to report sustainability policies than private firms.

a. Identify the population of interest to the researchers.
b. What method was used to collect the sample data?
c. Comment on the representativeness of the sample.
d. How will your answer to part **c** impact the validity of the inferences drawn from the study?

1.29 Inspection of highway bridges. All highway bridges in the United States are inspected periodically for structural deficiency by the Federal Highway Administration (FHWA). Data from the FHWA inspections are compiled into the National Bridge Inventory (NBI). Several of the nearly 100 variables maintained by the NBI are listed below. Classify each variable as quantitative or qualitative.

a. Length of maximum span (feet)
b. Number of vehicle lanes
c. Toll bridge (yes or no)
d. Average daily traffic
e. Condition of deck (good, fair, or poor)
f. Bypass or detour length (miles)
g. Route type (interstate, U.S., state, county, or city)

1.30 Structurally deficient highway bridges. Refer to Exercise 1.29. The NBI data were analyzed and the results made available at the FHWA Web site (www.fhwa.dot.gov). Using the FHWA inspection ratings, each of the 608,272 highway bridges in the United States was categorized as structurally deficient, functionally obsolete, or safe. About 13.5% of the bridges were found to be structurally deficient, while 3.5% were functionally obsolete.

a. What is the variable of interest to the researchers?
b. Is the variable of part a quantitative or qualitative?
c. Is the data set analyzed a population or a sample? Explain.
d. How did the NBI obtain the data for the study?

***1.31 Monitoring product quality.** The Wallace Company of Houston is a distributor of pipes, valves, and fittings to the refining, chemical, and petrochemical industries. The company was a recent winner of the Malcolm Baldrige National Quality Award. One of the steps the company takes to monitor the quality of its distribution process is to send out a survey twice a year to a subset of its current customers, asking the customers to rate the speed of deliveries, the accuracy of invoices, and the quality of the packaging of the products they have received from Wallace.

a. Describe the process studied.
b. Describe the variables of interest.
c. Describe the sample.
d. Describe the inferences of interest.
e. What are some of the factors that are likely to affect the reliability of the inferences?

1.32 Guilt in decision making. The effect of guilt emotion on how a decision maker focuses on the problem was investigated in the *Journal of Behavioral Decision Making* (January 2007). A total of 171 volunteer students participated in the experiment, where each was randomly assigned to one of three emotional states (guilt, anger, or neutral) through a reading/writing task. Immediately after the task, the students were presented with a decision problem (e.g., whether or not to spend money on repairing a very old car). The researchers found that a higher proportion of students in the guilty-state group chose to repair the car than those in the neutral-state and anger-state groups.

a. Identify the population, sample, and variables measured for this study.
b. Identify the data-collection method used.
c. What inference was made by the researcher?
d. In later chapters you will learn that the reliability of an inference is related to the size of the sample used. In addition to sample size, what factors might affect the reliability of the inference drawn in this study?

1.33 Accounting and Machiavellianism. *Behavioral Research in Accounting* (January 2008) published a study of Machiavellian traits in accountants. *Machiavellian* describes negative character traits that include manipulation, cunning, duplicity, deception, and bad faith. A questionnaire was administered to a random sample of 700 accounting alumni of a large southwestern university; however, due to nonresponse and incomplete answers, only 198 questionnaires could be analyzed. Several variables were measured, including age, gender, level of education, income, job satisfaction score, and Machiavellian ("Mach") rating score. The research findings suggest that Machiavellian behavior is not required to achieve success in the accounting profession.

a. What is the population of interest to the researcher?
b. What type of data (quantitative or qualitative) is produced by each of the variables measured?
c. Identify the sample.
d. Identify the data-collection method used.
e. What inference was made by the researcher?
f. How might the nonresponses impact the inference?

1.34 Can money spent on gifts buy love? Is the gift you purchased for that special someone really appreciated? This was the question of interest to business professors at Stanford University. Their research was published in the *Journal of Experimental Social Psychology* (Vol. 45, 2009). In one study, the researchers investigated the link between engagement ring price (dollars) and level of appreciation of the recipient (measured on a 7-point scale where 1 = "not at all" and 7 = "to a great extent"). Participants for the study were those who used a popular Web site for engaged couples. The Web site's directory was searched for those with "average" American names (e.g., "John Smith," "Sara Jones"). These individuals were then invited to participate in an online survey in exchange for a $10 gift certificate. Of the respondents, those who paid really high or really low prices for the ring were excluded, leaving a sample size of 33 respondents.

a. Identify the experimental units for this study.
b. What are the variables of interest? Are they quantitative or qualitative in nature?

c. Describe the population of interest.

d. Do you believe the sample of 33 respondents is representative of the population? Explain.

e. In a second designed study, the researchers investigated whether the link between gift price and level of appreciation is stronger for birthday gift-givers than for birthday gift-receivers. The participants were randomly assigned to play the role of gift-giver or gift-receiver. Assume that the sample consists of 50 individuals. Use a random number generator to randomly assign 25 individuals to play the gift-receiver role and 25 to play the gift-giver role.

NW

Applying the Concepts—Advanced

1.35 Bank corporate mergers. *Corporate merger* is a means through which one firm (the bidder) acquires control of the assets of another firm (the target). Recently, there was a frenzy of bank mergers in the United States, as the banking industry consolidated into more efficient and more competitive units.

a. Construct a brief questionnaire (two or three questions) that could be used to query a sample of bank presidents concerning their opinions of why the industry is consolidating and whether it will consolidate further.

b. Describe the population about which inferences could be made from the results of the survey.

c. Discuss the pros and cons of sending the questionnaire to all bank presidents versus a sample of 200.

1.36 Random-digit dialing. To ascertain the effectiveness of their advertising campaigns, firms frequently conduct telephone interviews with consumers using *random-digit dialing*. With this approach, a random number generator mechanically creates the sample of phone numbers to be called. Each digit in the phone number is randomly selected from the possible digits 0, 1, 2, . . . , 9. Use the procedure to generate five seven-digit telephone numbers whose first three digits (area code) are 373.

1.37 Current population survey. The employment status (employed or unemployed) of each individual in the U.S. workforce is a set of data that is of interest to economists, businesspeople, and sociologists. To obtain information about the employment status of the workforce, the U.S. Bureau of the Census conducts what is known as the *Current Population Survey*. Each month interviewers visit about 60,000 of the 128 million households in the United States and question the occupants 15 years of age or older about their employment status. Their responses enable the Bureau of the Census to *estimate* the percentage of people in the labor force who are unemployed (the *unemployment rate*).

a. Define the population of interest to the Census Bureau.

b. What variable is being measured? Is it quantitative or qualitative?

c. Is the problem of interest to the Census Bureau descriptive or inferential?

d. In order to monitor the rate of unemployment, it is essential to have a definition of *unemployed*. Different economists and even different countries define it in various ways. Develop your own definition of an "unemployed person." Your definition should answer such questions as: Are students on summer vacation unemployed? Are college professors who do not teach summer school unemployed? At what age are people considered to be eligible for the workforce? Are people who are out of work but not actively seeking a job unemployed?

***1.38 Monitoring the production of soft-drink cans.** The Wakefield plant of Coca-Cola and Schweppes Beverages Limited (CCSB) can produce 4,000 cans of soft drink per minute. The automated process consists of measuring and dispensing the raw ingredients into storage vessels to create the syrup, and then injecting the syrup, along with carbon dioxide, into the beverage cans. In order to monitor the subprocess that adds carbon dioxide to the cans, five filled cans are pulled off the line every 15 minutes, and the amount of carbon dioxide in each of these five cans is measured to determine whether the amounts are within prescribed limits.

NW

a. Describe the process studied.

b. Describe the variable of interest.

c. Describe the sample.

d. Describe the inference of interest.

e. *Brix* is a unit for measuring sugar concentration. If a technician is assigned the task of estimating the average brix level of all 240,000 cans of beverage stored in a warehouse near Wakefield, will the technician be examining a process or a population? Explain.

1.39 Sampling TV markets for a court case. A recent court case involved a claim of satellite television subscribers obtaining illegal access to local TV stations. The defendant (the satellite TV company) wanted to sample TV markets nationwide and determine the percentage of its subscribers in each sampled market who have illegal access to local TV stations. To do this, the defendant's expert witness drew a rectangular grid over the continental United States, with horizontal and vertical grid lines every .02 degrees of latitude and longitude, respectively. This created a total of 500 rows and 1,000 columns, or (500)(1,000) = 500,000 intersections. The plan was to randomly sample 900 intersection points and include the TV market at each intersection in the sample. Explain how you could use a random number generator to obtain a random sample of 900 intersections. Develop at least two plans: one that numbers the intersections from 1 to 500,000 prior to selection and another that selects the row and column of each sampled intersection (from the total of 500 rows and 1,000 columns).

Critical Thinking Challenge

1.40 *20/20* survey exposé. Refer to the "Statistics in Action" box of this chapter (p. 1). Recall that the popular primetime ABC television program *20/20* presented several misleading (and possibly unethical) surveys in a segment titled "Fact or Fiction?—Exposés of So-Called Surveys." The information reported from two of these surveys and several others is listed here (actual survey facts are provided in parentheses).

- *Quaker Oats study:* Eating oat bran is a cheap and easy way to reduce your cholesterol count. (Fact: Diet must consist of nothing but oat bran to achieve a slightly lower cholesterol count.)
- *March of Dimes report:* Domestic violence causes more birth defects than all medical issues combined. (Fact: No study—false report.)

- *American Association of University Women (AAUW) study:* Only 29% of high school girls are happy with themselves, compared to 66% of elementary school girls. (Fact: Of 3,000 high school girls, 29% responded, "Always true" to the statement "I am happy the way I am." Most answered, "Sort of true" and "Sometimes true.")
- *Food Research and Action Center study:* One in four American children under age 12 is hungry or at risk of hunger. (Fact: Based on responses to questions: "Do you ever cut the size of meals?" and "Do you ever eat less than you feel you should?" and "Did you ever rely on limited numbers of foods to feed your children because you were running out of money to buy food for a meal?")
- *McKinsey survey on the health reform act:* Thirty percent of employers would "definitely" or "probably" stop offering health coverage to their employees if the government-sponsored act is passed. (Fact: Employers were asked leading questions that made it seem logical for them to stop offering insurance. For example, respondents were told that the new health insurance exchanges would become "an easy, affordable way for individuals to obtain health insurance" outside the company. Then they were given examples of how little their workers would pay for this insurance. Only then were they asked how likely they would be to stop offering health insurance.)

a. Refer to the Quaker Oats study relating oat bran to cholesterol levels. Discuss why it is unethical to report the results as stated.

b. Consider the false March of Dimes report on domestic violence and birth defects. Discuss the type of data required to investigate the impact of domestic violence on birth defects. What data-collection method would you recommend?

c. Refer to the AAUW study of self-esteem of high school girls. Explain why the results of the study are likely to be misleading. What data might be appropriate for assessing the self-esteem of high school girls?

d. Refer to the Food Research and Action Center study of hunger in America. Explain why the results of the study are likely to be misleading. What data would provide insight into the proportion of hungry American children?

e. Refer to the McKinsey survey on the health reform act. Explain what a "leading question" is and why it might produce responses that bias the results.

ACTIVITY 1.1 *Keep the Change:* Collecting Data

Bank of America has a savings program called *Keep the Change.* Each time a customer enrolled in the program uses his or her debit card to make a purchase, the difference between the purchase total and the next higher dollar amount is transferred from the customer's checking account to a savings account. For example, if you were enrolled in the program and used your debit card to purchase a latte for $3.75, then $0.25 would be transferred from your checking to your savings account. For the first 90 days that a customer is enrolled in the program, Bank of America matches the amounts transferred up to $250. In this and subsequent activities, we will investigate the potential benefit to the customer and cost to the bank.

1. Simulate the program by keeping track of all purchases that you make during one week that could be made with a debit card, even if you use a different form of payment. For each purchase, record both the purchase total and the amount that would be transferred from checking to savings with the *Keep the Change* program.
2. You now have two sets of data: *Purchase Totals* and *Amounts Transferred*. Both sets contain quantitative data. For each data set, identify the corresponding naturally occurring numerical scale. Explain why each set has an obvious lower bound but only one set has a definite upper bound.
3. Find the total of the amounts transferred for the one-week period. Because 90 days is approximately 13 weeks, multiply the total by 13 to estimate how much the bank would have to match during the first 90 days. Form a third data set, *Bank Matching*, by collecting the 90-day estimates of all the students in your class. Identify the naturally occurring scale, including bounds, for this set of data.

Keep the data sets from this activity for use in other activities. We suggest you save the data using statistical software (e.g., Minitab) or a graphing calculator.

ACTIVITY 1.2 Identifying Misleading Statistics

In the *Statistics in Action* feature at the beginning of this chapter, several examples of false or misleading statistics were discussed. Claims such as *One in four American children under age 12 is hungry or at risk of hunger* are often used to persuade the public or the government to donate or allocate more money to charitable groups that feed the poor. Researchers sometimes claim a relationship exists between two seemingly unrelated quantities such as a CEO's golf handicap and the company's stock performance; such relationships are often weak at best and of little practical importance. Read the *Statistics in Action* and *Statistics in Action Revisited* features in this chapter before completing this activity.

1. Look for an article in a newspaper or on the Internet in which a large proportion or percentage of a population is purported to be "at risk" of some calamity, as in the childhood hunger example. Does the article cite a source or provide any information to support the proportion or

(*continued*)

percentage reported? Is the goal of the article to persuade some individual or group to take some action? If so, what action is being requested? Do you believe that the writer of the article may have some motive for exaggerating the problem? If so, give some possible motives.

2. Look for another article in which a relationship between two seemingly unrelated quantities is purported to exist, as in the CEO golf handicap and stock performance study. Select an article that contains some information on how the data were collected. Identify the target population and the data-collection method. Based on what is presented in the article, do you believe that the data are representative of the population? Explain. Is the purported relationship of any practical interest? Explain.

References

Careers in Statistics, American Statistical Association, 2011 (www.amstat.org).

Cochran, W. G. *Sampling Techniques*, 3rd ed. New York: Wiley, 1977.

Deming, W. E. *Sample Design in Business Research.* New York: Wiley, 1963.

Dillman, D. A., Smyth, J. D., and Christian, L. M. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. New York: Wiley, 2008.

Ethical Guidelines for Statistical Practice, American Statistical Association, 1999 (www.amstat.org).

Hahn, G. J. and Doganaksoy, N. *The Role of Statistics in Business and Industry*. New York: Wiley, 2008.

Huff, D. *How to Lie with Statistics*. New York: Norton, 1982 (paperback 1993).

Hoerl, R. and Snee, R. *Statistical Thinking: Improving Business Performance*. Boston: Duxbury, 2002.

Kish, L. *Survey Sampling.* New York: Wiley, 1965 (paperback, 1995).

Peck, R., Casella, G., Cobb, G., Hoerl, R., Nolan, D., Starbuck, R., and Stern, H. *Statistics: A Guide to the Unknown*, 4th ed. Cengage Learning, 2005.

Scheaffer, R., Mendenhall, W., and Ott, R. L. *Elementary Survey Sampling*, 6th ed. Boston: Duxbury, 2005.

What Is a Survey? American Statistical Association (F. Scheuren, editor), 2nd ed., 2005 (www.amstat.org).

USING TECHNOLOGY

Technology images shown here are taken from Minitab 19, XLSTAT 2019 and StatCrunch 3.0.

Minitab: Accessing and Listing Data

When you start a Minitab session, you will see a screen similar to Figure 1.M.1. The bottom portion of the screen is an empty spreadsheet—called a Minitab worksheet—with columns representing variables and rows representing observations (or cases). The very top of the screen is the Minitab main menu bar, with buttons for the different functions and procedures available in Minitab. Once you have entered data into the spreadsheet, you can analyze the data by clicking the appropriate menu buttons. The results will appear in the window above the worksheet.

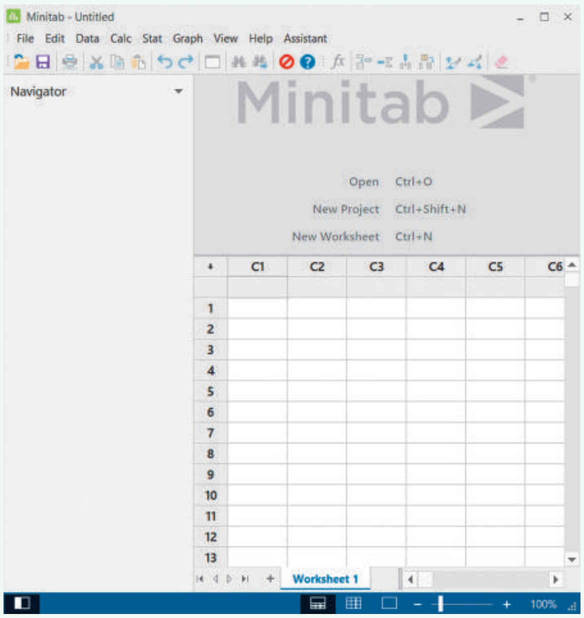


Figure 1.M.1 Initial screen viewed by the Minitab user

Entering Data

Create a Minitab data file by entering data directly into the worksheet. Figure 1.M.2 shows data entered for a variable called "GPA." Name the variables (columns) by typing in the name of each variable in the box below the column number.

↓	C1	C2
	GPA	
1	2.66	
2	3.50	
3	3.91	
4	2.85	
5	3.04	

Figure 1.M.2 Data entered into the Minitab worksheet

Opening a Minitab Data File

If the data have been previously saved as a Minitab (.mtw) file, access the data as follows.

Step 1 Click the "File" button on the menu bar, and then click "Open" as shown in Figure 1.M.3. A dialog box similar to Figure 1.M.4 will appear.

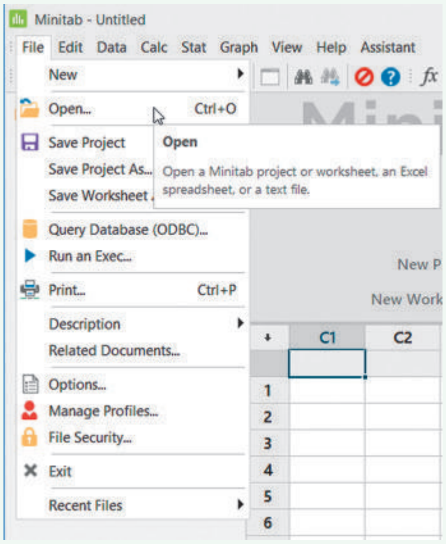


Figure 1.M.3 Options for opening a Minitab data file

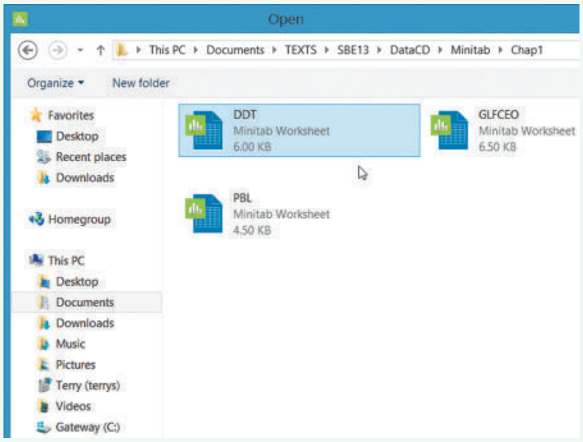


Figure 1.M.4 Selecting the Minitab data file to open

Step 2 Specify the location (folder) that contains the data, click on the Minitab data file, and then click "Open" (see Figure 1.M.4). The data will appear in the Minitab worksheet, as shown in Figure 1.M.5 at bottom of the page.

Accessing Data from an Excel File

Step 1 Click the "File" button on the menu bar, and then click "Open" as shown in Figure 1.M.3.

Step 2 Specify the location (folder) that contains the Excel data file and the file type (e.g., .xls), and then click on the file name, and "Open." This reveals the dialog box shown in Figure 1.M.6.

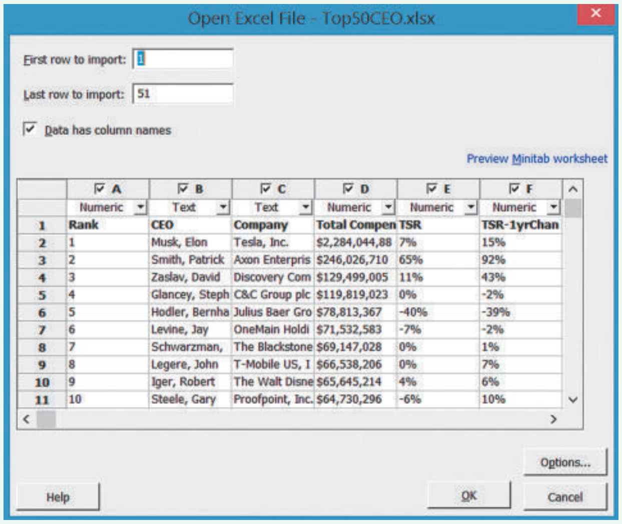


Figure 1.M.6 Open Excel file dialog box

Step 3 If the Excel spreadsheet has column names, check the appropriate box on the screen. Then click "OK." The data will appear in the MINITAB worksheet.

Reminder: If the Excel spreadsheet does not have column names, the variables (columns) can be named by typing in the name of each variable in the box under the column number.

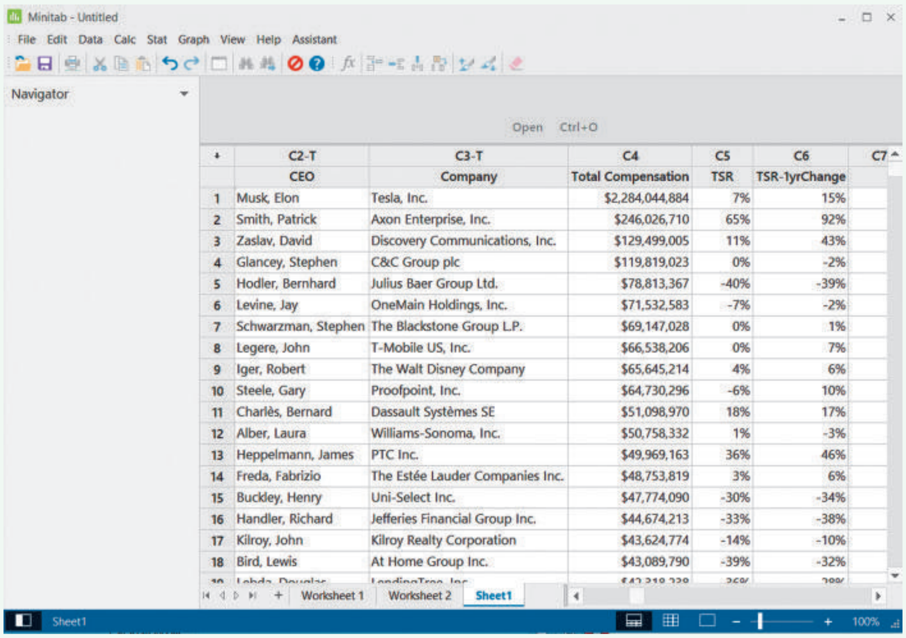


Figure 1.M.5 The Minitab worksheet showing the opened Minitab data file

Listing (Printing) Data

Step 1 Click on the "Data" button on the Minitab main menu bar, and then click on "Display Data." (See Figure 1.M.7.) The resulting menu, or dialog box, appears as in Figure 1.M.8.

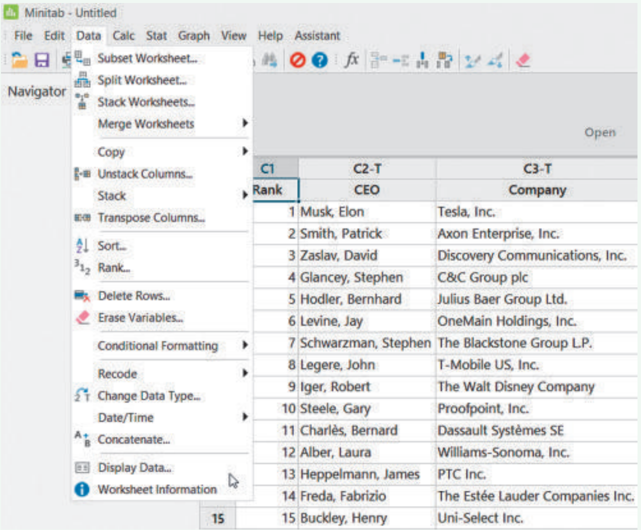


Figure 1.M.7 Minitab options for displaying data

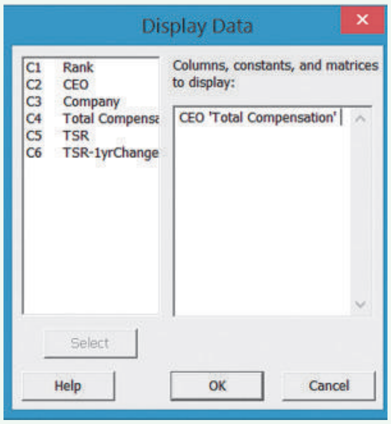


Figure 1.M.8 Minitab Display Data dialog box

Step 2 Enter the names of the variables you want to print in the "Columns, constants, and matrices to display" box (you can do this by simply double clicking on the variables), and then click "OK." The printout will show up on your Minitab session screen.

Minitab: Generating a Random Sample

Step 1 Click on the "Calc" button on the Minitab menu bar and then click on "Random Data," and finally, click on "Sample From Columns," as shown in Figure 1.M.9. The resulting dialog box appears as shown in Figure 1.M.10.

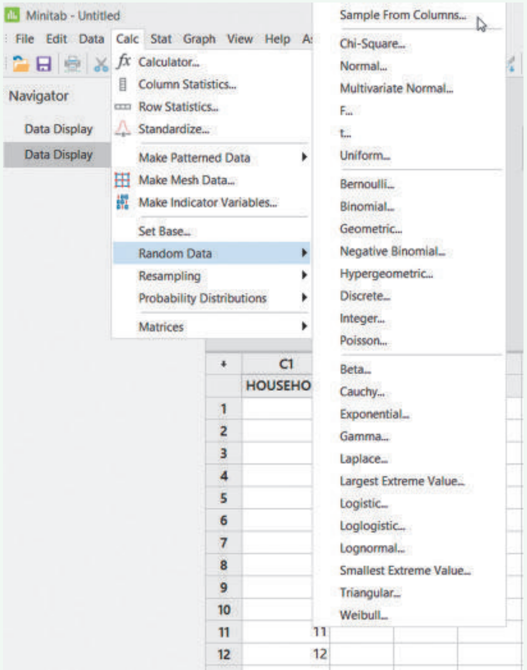


Figure 1.M.9 Minitab menu options for sampling from a data set

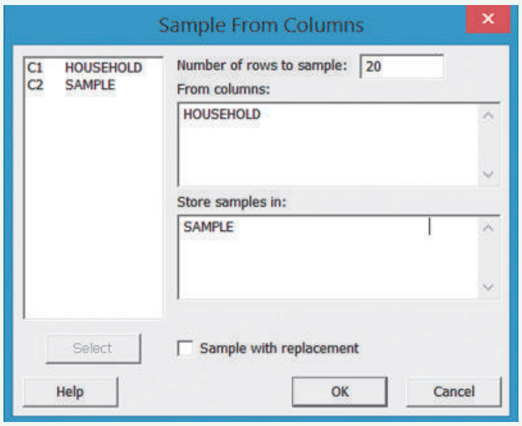


Figure 1.M.10 Minitab options for selecting a random sample from worksheet columns

Step 2 Specify the sample size (i.e., number of rows), the variable(s) to be sampled, and the column(s) where you want to save the sample.

Step 3 Click "OK" and the Minitab worksheet will reappear with the values of the variable for the selected (sampled) cases in the column specified.

In Minitab, you can also generate a sample of case numbers (e.g., a sample of cases from a population of cases numbered from 1 to 500).

Step 1 From the Minitab menu, click on the "Calc" button and then click on "Random Data," and finally, click on the "Integer" option (see Figure 1.M.9).

Step 2 In the resulting dialog box (shown in Figure 1.M.11), specify the number of cases (rows, i.e., the sample size), and the column where the case numbers selected will be stored. Also, specify the total number of observations in the population in the "Maximum value" box.

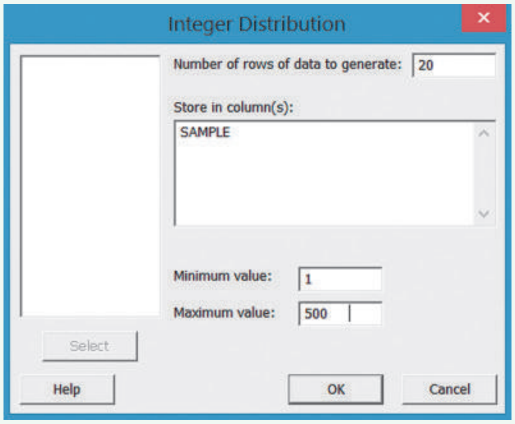


Figure 1.M.11 Minitab options for selecting a random sample of cases

Step 3 Click "OK" and the Minitab worksheet will reappear with the case numbers for the selected (sampled) cases in the column specified.

[*Note:* If you want the option of generating the same (identical) sample multiple times from the data set, then first click on the "Set Base" option shown in Figure 1.M.9. Specify an integer in the resulting dialog box. If you always select the same integer. Minitab will select the same sample when you choose the random sampling options.]

XLSTAT: Accessing and Listing Data

When you open XLSTAT, you will see a screen similar to Figure 1.E.1. The majority of the screen window is a spreadsheet—called an Excel workbook—with columns (labeled A, B, C, etc.) representing variables, and rows representing observations (or cases). The very top of the screen is the Excel main menu bar, with buttons for the different functions and procedures available in Excel. XLSTAT will be one of the menu options. Once you have entered data into the spreadsheet, you can analyze the data by clicking the appropriate XLSTAT menu buttons. The results will appear in a new workbook.

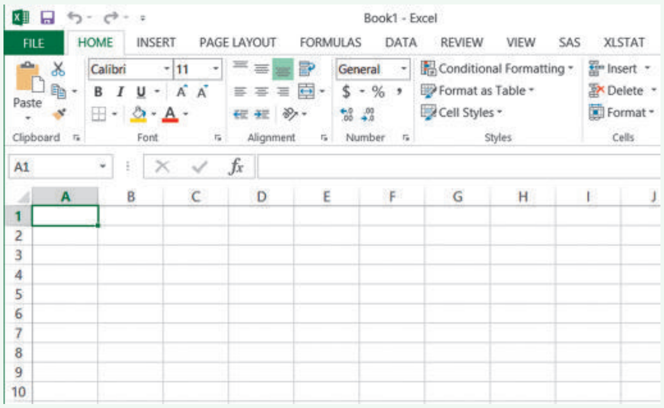


Figure 1.E.1 Initial screen viewed by XLSTAT user

Entering Data

Enter data directly into the appropriate row and column of the spreadsheet. Figure 1.E.2 shows data entered in the first (A) column. Optionally, you can add names for the variables (columns) in the first row of the workbook.

	A	B
1	GPA	
2	2.66	
3	3.91	
4	3.08	
5	2.74	
6	3.22	
7		

Figure 1.E.2 Data entered into the Excel workbook

Opening an Excel File

If the data have been previously saved as an Excel (.xls) file, access the data as follows.

Step 1 Click "File" at the far left of the menu bar, and then click "Open," as shown in Figure 1.E.3.

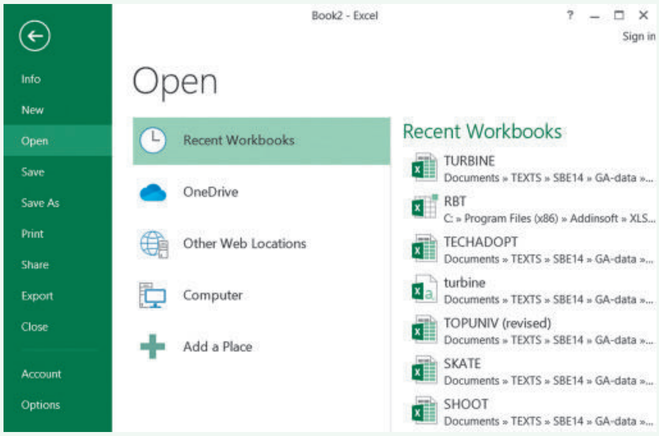


Figure 1.E.3 Options for opening an Excel file

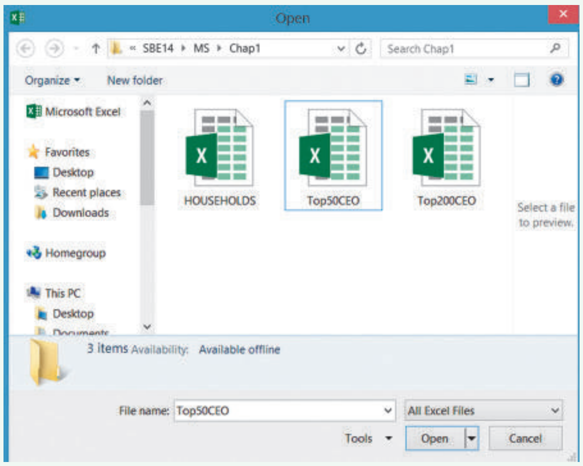


Figure 1.E.4 Selecting the Excel file to open