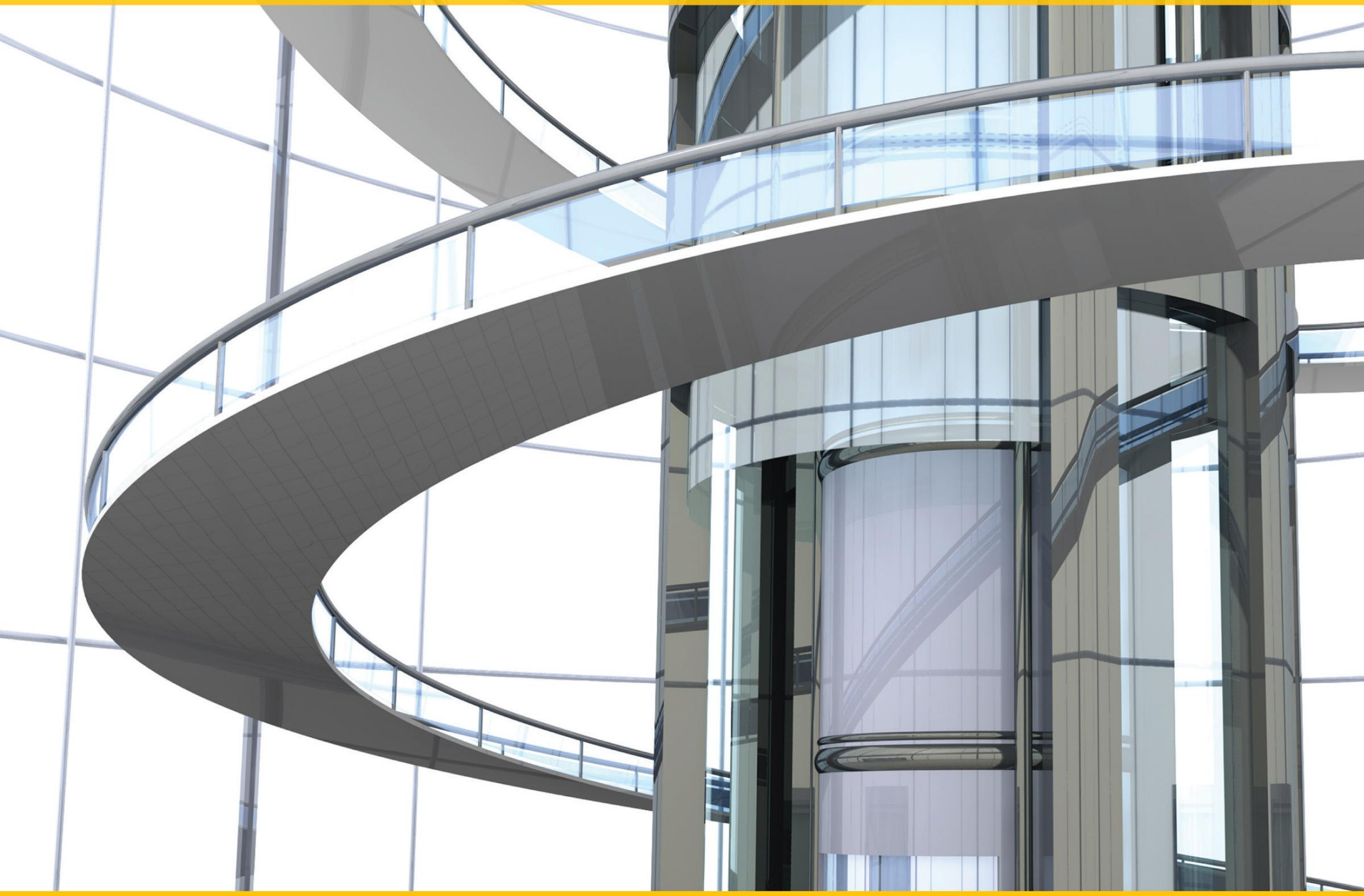
Anderson * Sweeney * Williams * Camm * Cochran



Essentials of Modern Business Statistics 7e

with Microsoft® Office Excel®

Want to turn your C into an A? Obviously, right?

But the right way to go about it isn't always so obvious. Go digital to get the grades. MindTap's customizable study tools and eTextbook give you everything you need all in one place.

Engage with your course content, enjoy the flexibility of studying anytime and anywhere, stay connected to assignment due dates and instructor notifications with the MindTap Mobile app...

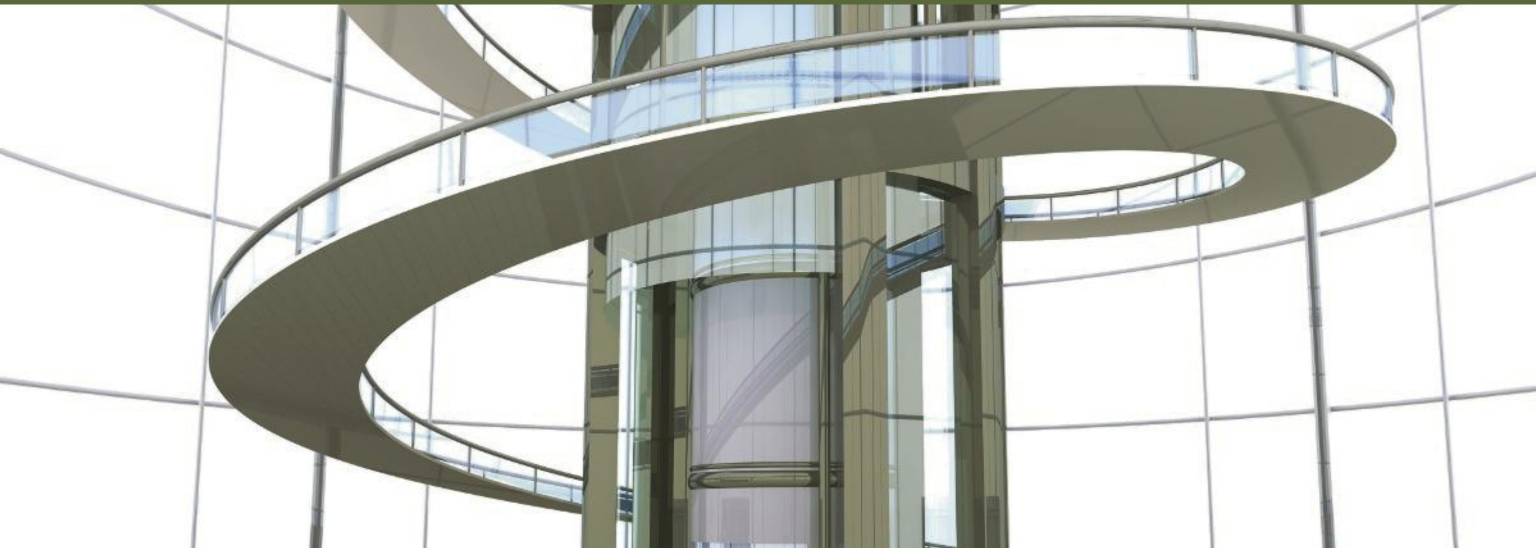
*and most of all...*EARN BETTER GRADES.

TO GET STARTED VISIT
WWW.CENGAGE.COM/STUDENTS/MINDTAP

CENGAGE Learning® | MindTap®

Microsoft® Office Excel® Functions

Function	Description
AVERAGE	Returns the arithmetic mean of a range its arguments.
BINOM.DIST	Returns the individual term binomial distribution probability.
CHISQ.DIST	Returns a probability from the chi-squared distribution.
CHISQ.DIST.RT	Returns the one-tailed probability of the chi-squared distribution.
CHISQ.INV	Returns the inverse of the left-tailed probability of the chi-squared distribution.
CHISQ.TEST	Returns the value from the chi-squared distribution for the statistic and the degrees of freedom.
CONFIDENCE.NORM	Returns the confidence interval for a population mean using the normal distribution.
CORREL	Returns the correlation coefficient between two data sets.
COUNT	Returns the number of cells in the range that contain numbers.
COUNTA	Returns the number of non-blank cells in the range.
COUNTIF	Returns the number of cells in a range that meet the specified criterion.
COVARIANCE.S	Returns the sample covariance.
EXPON.DIST	Returns a probability from the exponential distribution.
F.DIST.RT	Returns the right-tailed probability from the F distribution.
GEOMEAN	Returns the geometric mean of a range of cells.
HYPGEOM.DIST	Returns a probability from the hypergeometric distribution.
MAX	Returns the maximum value of the values in a range of cells.
MMEDIAN	Returns the median value of the values in a range of cells.
MIN	Returns the minimum value of the values in a range of cells.
MODE.SNGL	Returns the most-frequently occurring value in a range of cells.
NORM.S.DIST	Returns a probability from a standard normal distribution.
NORM.S.INV	Inverse of the standard normal distribution.
PERCENTILE.EXC	Returns the specified percentile of the values in a range of cells.
POISSON.DIST	Returns a probability from the poisson distribution.
POWER	Returns the result of a number raised to a power.
QUARTILE.EXC	Returns the specified quartile of the values in a range of cells.
RAND	Returns a real number from the uniform distribution between 0 and 1.
SQRT	Returns the positive square root of its argument.
STDEV.S	Returns the sample standard deviation of the values in a range of cells.
SUM	Returns the sum of the values in a range of cells.
SUMPRODUCT	Returns the sum of the products of the paired elements of the values in two ranges of cells.
T.DIST	Returns a left-tailed probability of the t distribution.
T.INV.2T	Returns the two-tailed inverse of the student's t-distribution.
VAR.S	Returns the sample variance of the values in a range of cells.



Essentials of Modern Business Statistics 7e

with Microsoft® Office Excel®

David R. Anderson
University of Cincinnati

Dennis J. Sweeney
University of Cincinnati

Thomas A. Williams
Rochester Institute
of Technology

Jeffrey D. Camm
Wake Forest University

James J. Cochran
University of Alabama

CENGAGE Learning®

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

CENGAGE
Learning®

Essentials of Modern Business Statistics, Seventh Edition

David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran

Senior Vice President, General Manager, Social Science, Business and Humanities: Erin Joyner

Product Director: Michael Schenk

Product Team Manager: Joe Sabatino

Sr. Product Manager: Aaron Arnsparger

Content Developer: Anne Merrill

Product Assistant: Renee Schnee

Digital Content Designer: Brandon Foltz

Sr. Marketing Manager: Nathan Anderson

Sr. Content Project Manager: Colleen Farmer

Manufacturing Planner: Ron Montgomery

Production Service: MPS Limited

Sr. Art Director: Michelle Kunkler

Internal Designer: Beckmeyer Design

Cover Designer: Beckmeyer Design

Cover Image: Spectral-Design/ Shutterstock.com

Intellectual Property:
Analyst: Brittani Morgan
Project Manager: Nicholas Barrows

For product information and technology assistance, contact us at
Cengage Learning Customer & Sales Support, 1-800-354-9706

For permission to use material from this text or product,
submit all requests online at **www.cengage.com/permissions**
Further permissions questions can be emailed to
permissionrequest@cengage.com

Library of Congress Control Number: 2017930365

Package ISBN: 978-1-337-29829-2

Book only ISBN: 978-1-337-29835-3

Cengage Learning
20 Channel Center Street
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with employees residing in nearly 40 different countries and sales in more than 125 countries around the world. Find your local representative at **www.cengage.com**.

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage Learning Solutions, visit **www.cengage.com**

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com**

Printed in the United States of America
Print Number: 01 Print Year: 2017

Brief Contents

Contents

Chapter 2 Descriptive Statistics: Tabular and Graphical Displays 35

Chapter 7 Sampling and Sampling Distributions 317

Chapter 8 Interval Estimation 363

Chapter 9 Hypothesis Tests 405

Preface

This text is the seventh edition of *Essentials of Modern Business Statistics with Microsoft® Office Excel®*. With this edition we welcome two eminent scholars to our author team: Jeffrey D. Camm of Wake Forest University and James J. Cochran of the University of Alabama. Both Jeff and Jim are accomplished teachers, researchers, and practitioners in the fields of statistics and business analytics. Jim is a fellow of the American Statistical Association. You can read more about their accomplishments in the About the Authors section that follows this preface. We believe that the addition of Jeff and Jim as our coauthors will both maintain and improve the effectiveness of *Essentials of Modern Business Statistics with Microsoft Office Excel*.

The purpose of *Essentials of Modern Business Statistics with Microsoft® Office Excel®* is to give students, primarily those in the fields of business administration and economics, a conceptual introduction to the field of statistics and its many applications. The text is applications oriented and written with the needs of the nonmathematician in mind; the mathematical prerequisite is knowledge of algebra.

Applications of data analysis and statistical methodology are an integral part of the organization and presentation of the text material. The discussion and development of each technique is presented in an applications setting, with the statistical results providing insights for decision making and solutions to applied problems.

Although the book is applications oriented, we have taken care to provide sound methodological development and to use notation that is generally accepted for the topic being covered. Hence, students will find that this text provides good preparation for the study of more advanced statistical material. A bibliography to guide further study is included as an appendix.

Use of Microsoft Excel for Statistical Analysis

Essentials of Modern Business Statistics with Microsoft® Office Excel® is first and foremost a statistics textbook that emphasizes statistical concepts and applications. But since most practical problems are too large to be solved using hand calculations, some type of statistical software package is required to solve these problems. There are several excellent statistical packages available today; however, because most students and potential employers value spreadsheet experience, many schools now use a spreadsheet package in their statistics courses. Microsoft Excel is the most widely used spreadsheet package in business as well as in colleges and universities. We have written *Essentials of Modern Business Statistics with Microsoft® Office Excel®* especially for statistics courses in which Microsoft Excel is used as the software package.

Excel has been integrated within each of the chapters and plays an integral part in providing an application orientation. Although we assume that readers using this text are familiar with Excel basics such as selecting cells, entering formulas, copying, and so on, we do not assume that readers are familiar with Excel 2016 or Excel's tools for statistical analysis. As a result, we have included Appendix E, which provides an introduction to Excel 2016 and tools for statistical analysis.

Throughout the text the discussion of using Excel to perform a statistical procedure appears in a subsection immediately following the discussion of the statistical procedure. We believe that this style enables us to fully integrate the use of Excel throughout the text, but still maintain the primary emphasis on the statistical methodology being discussed. In each of these subsections, we provide a standard format for using Excel for statistical analysis. There are four primary tasks: Enter/Access Data, Enter Functions and Formulas, Apply Tools, and Editing Options. The Editing Options task is new with this edition. It primarily involves how to edit Excel output so that it is more suitable for presentations to users. We believe a consistent framework for applying Excel helps users to focus on the statistical methodology without getting bogged down in the details of using Excel.

In presenting worksheet figures, we often use a nested approach in which the worksheet shown in the background of the figure displays the formulas and the worksheet shown in the foreground shows the values computed using the formulas. Different colors and shades of colors are used to differentiate worksheet cells containing data, highlight cells containing Excel functions and formulas, and highlight material printed by Excel as a result of using one or more data analysis tools.

Changes in the Seventh Edition

We appreciate the acceptance and positive response to the previous editions of *Essentials of Modern Business Statistics with Microsoft® Office Excel®*. Accordingly, in making modifications for this new edition, we have maintained the presentation style and readability of those editions. The significant changes in the new edition are summarized here.

Users of the previous edition will notice that the chapters offered and topics covered in this edition differ from previous editions. While the topical coverage of the first nine chapters remains the same, the organization and coverage in some of the later chapters have expanded. We have eliminated the coverage of the advanced topic of Time Series and Quality Control in favor of the expanded coverage in Chapters 10, 11, 12 and 13. Chapter 10 now provides coverage of inferences of means and proportions for two populations, and chapter 11 is focused on inferences about population variances. Chapter 12 is a discussion of comparing multiple proportions, tests of independence and goodness of fit and chapter 13 covers experimental design and ANOVA. We believe you will find the expanded coverage in these chapters useful in your classes. Coverage of regression is now in chapters 14 and 15. These two chapters are revisions of the regression chapters from the 6th edition. In additions to these changes, we made the following revisions:

- **Microsoft Excel 2016.** Step-by-step instructions and screen captures show how to use the latest version of Excel to implement statistical procedures.
- **Data and Statistics—Chapter 1.** We have expanded our section on data mining to include a discussion of big data. We have added a new section on analytics. We have also placed greater emphasis on the distinction between observed and experimental data.
- **Descriptive Statistics: Tabular and Graphical Displays—Chapter 2.** Microsoft Excel now has the capability of creating box plots and comparative box plots. We have added to this chapter instruction on how to use this very useful new feature.
- **Interval Estimation—Chapter 8.** We have added a new section on the implications of big data (large data sets) on the interpretation of confidence intervals and importantly, the difference between statistical and practical significance.
- **Hypothesis Tests—Chapter 9.** Similar to our addition to Chapter 8, we have added a new section on the implications of big data (large data sets) on the interpretation of hypothesis tests and the difference between statistical and practical significance.
- **Simple Linear Regression—Chapter 14.** Similar to our addition to Chapter 8, we have added a new section on the implications of big data (large data sets) on

the interpretation of hypothesis tests in simple linear regression and the difference between statistical and practical significance.

- **New Case Problems.** We have added thirteen new case problems to this edition. The new case problems appear in the chapters on descriptive statistics and regression analysis. The case problems in the text provide students with the opportunity to analyze somewhat larger data sets and prepare managerial reports based on the results of their analysis.
- **New Examples and Exercises Based on Real Data.** We have added approximately 126 new examples and exercises based on real data and recently referenced sources of statistical information. Using data obtained from various data collection organizations, websites, and other sources such as *The Wall Street Journal*, *USA Today*, *Fortune*, and *Barron's*, we have drawn upon actual studies to develop explanations and to create exercises that demonstrate many uses of statistics in business and economics. We believe the use of real data helps generate more student interest in the material and enables the student to learn about both the statistical methodology and its application.
- **Updated and Improved End-of-Chapter Solutions and Solutions Manual.** Partnering with accomplished instructor Dawn Bulriss at Maricopa Community Colleges, we took a deep audit of the solutions manual. Every question and solution was reviewed and reworked, as necessary. The solutions now contain additional detail: improved rounding instructions; expanded explanations with a student-focus; and alternative answers using Excel and a statistical calculator. We believe this thorough review will enhance both the instructor and student learning experience in this digital age.

Features and Pedagogy

Authors Anderson, Sweeney, Williams, Camm, and Cochran have continued many of the features that appeared in previous editions.

Methods Exercises and Applications Exercises

The end-of-section exercises are split into two parts, Methods and Applications. The Methods exercises require students to use the formulas and make the necessary computations. The Applications exercises require students to use the chapter material in real-world situations. Thus, students first focus on the computational "nuts and bolts" and then move on to the subtleties of statistical application and interpretation.

Self-Test Exercises

Certain exercises are identified as self-test exercises. Completely worked-out solutions for those exercises are provided in Appendix D in the Student Resources online. Students can attempt the self-test exercises and immediately check the solution to evaluate their understanding of the concepts presented in the chapter.

Margin Annotations and Notes and Comments

Margin annotations that highlight key points and provide additional insights for the students are a key feature of this text. These annotations are designed to provide emphasis and enhance understanding of the terms and concepts being presented in the text.

At the end of many sections, we provide Notes and Comments designed to give the student additional insights about the statistical methodology and its application. Notes and Comments include warnings about or limitations of the methodology, recommendations for application, brief descriptions of additional technical considerations, and other matters.

Data Files Accompany the Text

Approximately 220 data files are available on the website that accompanies this text. The data sets are available in Excel 2016 format. DATAfile logos are used in the text to identify the data sets that are available on the website. Data sets for all case problems as well as data sets for larger exercises are included.

MindTap

MindTap, featuring all new Excel Online integration powered by Microsoft, is a complete digital solution for the business statistics course. It has enhancements that take students from learning basic statistical concepts to actively engaging in critical thinking applications, while learning valuable software skills for their future careers.

MindTap is a customizable digital course solution that includes an interactive eBook, autograded, algorithmic exercises from the textbook, Adaptive Test Prep, as well as interactive visualizations. All of these materials offer students better access to understand the materials within the course. For more information on MindTap, please contact your Cengage representative.

For Students

Online resources are available to help the student work more efficiently. The resources can be accessed through **www.cengagebrain.com**.

For Instructors

Instructor resources are available to adopters on the Instructor Companion Site, which can be found and accessed at **www.cengage.com**, including:

- **Solutions Manual:** The Solutions Manual, prepared by the authors, includes solutions for all problems in the text. It is available online as well as print.
- **Solutions to Case Problems:** These are also prepared by the authors and contain solutions to all case problems presented in the text.
- **PowerPoint Presentation Slides:** The presentation slides contain a teaching outline that incorporates figures to complement instructor lectures.
- **Test Bank:** Cengage Learning Testing Powered by Cognero is a flexible, online system that allows you to:
  - author, edit, and manage test bank content from multiple Cengage Learning solutions,
  - create multiple test versions in an instant, and
  - deliver tests from your LMS, your classroom, or wherever you want. The Test Bank is also available in Microsoft Word.

Acknowledgments

A special thanks goes to our associates from business and industry who supplied the Statistics in Practice features. We recognize them individually by a credit line in each of the articles. We are also indebted to our product manager, Aaron Arnsparger; our content developer, Anne Merrill; our content project manager, Colleen Farmer; our project manager at MPS Limited, Gaurav Prabhu; digital content designer, Brandon Foltz; and others at Cengage for their editorial counsel and support during the preparation of this text.

We would like to acknowledge the work of our reviewers, who provided comments and suggestions of ways to continue to improve our text. Thanks to:

James Bang, Virginia Military Institute
Robert J. Banis, University of Missouri–St. Louis
Timothy M. Bergquist, Northwest Christian College
Gary Black, University of Southern Indiana
William Bleuel, Pepperdine University
Derrick Boone, Wake Forest University
Lawrence J. Bos, Cornerstone University
Dawn Bulriss, Maricopa Community Colleges
Joseph Cavanaugh, Wright State University–Lake Campus
Sheng-Kai Chang, Wayne State University
Robert Christopherson, SUNY-Plattsburgh
Michael Clark, University of Baltimore
Robert D. Collins, Marquette University
Ivona Contardo, Stellenbosch University
Sean Eom, Southeast Missouri State University
Samo Ghosh, Albright College
Philip A. Gibbs, Washington & Lee University
Daniel L. Gilbert, Tennessee Wesleyan College
Michael Gorman, University of Dayton
Erick Hofacker, University of Wisconsin, River Falls
David Juriga, St. Louis Community College
William Kasperski, Madonna University
Kuldeep Kumar, Bond Business School
Tenpao Lee, Niagara University
Ying Liao, Meredith College
Daniel Light, Northwest State College
Ralph Maliszewski, Waynesburg University
Saverio Manago, Salem State University
Patricia A. Mullins, University of Wisconsin–Madison
Jack Muryn, Cardinal Stritch University
Anthony Narsing, Macon State College
Robert M. Nauss, University of Missouri–St. Louis
Elizabeth L. Rankin, Centenary College of Louisiana
Surekha Rao, Indiana University, Northwest
Jim Robison, Sonoma State University
Farhad Saboori, Albright College
Susan Sandblom, Scottsdale Community College
Ahmad Saranjam, Bridgewater State University
Jeff Sarbaum, University of North Carolina at Greensboro
Robert Scott, Monmouth University

Toni Somers, Wayne State University
Jordan H. Stein, University of Arizona
Bruce Thompson, Milwaukee School of Engineering
Ahmad Vessal, California State University, Northridge
Dave Vinson, Pellissippi State
Daniel B. Widdis, Naval Postgraduate School
Peter G. Wagner, University of Dayton
Sheng-Ping Yang, Black Hills State University

We would like to recognize the following individuals, who have helped us in the past and continue to influence our writing.

Glen Archibald, University of Mississippi
Darl Bien, University of Denver
Thomas W. Bolland, Ohio University
Mike Bourke, Houston Baptist University
Peter Bryant, University of Colorado
Terri L. Byczkowski, University of Cincinnati
Robert Carver, Stonehill College
Ying Chien, University of Scranton
Robert Cochran, University of Wyoming
Murray Côté, University of Florida
David W. Cravens, Texas Christian University
Eddine Dahel, Monterey Institute of International Studies
Tom Dahlstrom, Eastern College
Terry Dielman, Texas Christian University
Joan Donohue, University of South Carolina
Jianjun Du, University of Houston–Victoria
Thomas J. Dudley, Pepperdine University
Swarna Dutt, University of West Georgia
Ronald Ehresman, Baldwin-Wallace College
Mohammed A. El-Saidi, Ferris State University
Robert Escudero, Pepperdine University
Stacy Everly, Delaware County Community College
Soheila Kahkashan Fardanesh, Towson University
Nicholas Farnum, California State University–Fullerton
Abe Feinberg, California State University, Northridge
Michael Ford, Rochester Institute of Technology
Phil Fry, Boise State University
V. Daniel Guide, Duquesne University
Paul Guy, California State University–Chico
Charles Harrington, University of Southern Indiana
Carl H. Hess, Marymount University
Woodrow W. Hughes, Jr., Converse College
Alan Humphrey, University of Rhode Island
Ann Hussein, Philadelphia College of Textiles and Science
Ben Isselhardt, Rochester Institute of Technology
Jeffery Jarrett, University of Rhode Island
Barry Kadets, Bryant College
Homayoun Khamooshi, George Washington University
Kenneth Klassen, California State University Northridge
David Krueger, St. Cloud State University
June Lapidus, Roosevelt University

Martin S. Levy, University of Cincinnati
Daniel M. Light, Northwest State College
Ka-sing Man, Georgetown University
Don Marx, University of Alaska, Anchorage
Tom McCullough, University of California–Berkeley
Timothy McDaniel, Buena Vista University
Mario Miranda, The Ohio State University
Barry J. Monk, Macon State College
Mitchell Muesham, Sam Houston State University
Richard O'Connell, Miami University of Ohio
Alan Olinsky, Bryant College
Lynne Pastor, Carnegie Mellon University
Von Roderick Plessner, Northwest State University
Robert D. Potter, University of Central Florida
Tom Pray, Rochester Institute of Technology
Harold Rahmlow, St. Joseph's University
Derrick Reagle, Fordham University
Avuthu Rami Reddy, University of Wisconsin–Platteville
Tom Ryan, Case Western Reserve University
Ahmad Saranjam, Bridgewater State College
Bill Seaver, University of Tennessee
Alan Smith, Robert Morris College
William Struning, Seton Hall University
Ahmad Syamil, Arkansas State University
David Tufte, University of New Orleans
Jack Vaughn, University of Texas–El Paso
Elizabeth Wark, Springfield College
Ari Wijetunga, Morehead State University
Nancy A. Williams, Loyola College in Maryland
J. E. Willis, Louisiana State University
Larry Woodward, University of Mary Hardin–Baylor
Mustafa Yilmaz, Northeastern University

David R. Anderson
Dennis J. Sweeney
Thomas A. Williams
Jeffrey D. Camm
James J. Cochran

CHAPTER 1

Data and Statistics

CONTENTS

STATISTICS *in* PRACTICE

BLOOMBERG BUSINESSWEEK*

NEW YORK, NEW YORK

With a global circulation of more than 1 million, *Bloomberg Businessweek* is one of the most widely read business magazines in the world. Bloomberg's 1700 reporters in 145 service bureaus around the world enable *Bloomberg Businessweek* to deliver a variety of articles of interest to the global business and economic community. Along with feature articles on current topics, the magazine contains articles on international business, economic analysis, information processing, and science and technology. Information in the feature articles and the regular sections helps readers stay abreast of current developments and assess the impact of those developments on business and economic conditions.

Most issues of *Bloomberg Businessweek,* formerly *BusinessWeek,* provide an in-depth report on a topic of current interest. Often, the in-depth reports contain statistical facts and summaries that help the reader understand the business and economic information. Examples of articles and reports include the impact of businesses moving important work to cloud computing, the crisis facing the U.S. Postal Service, and why the debt crisis is even worse than we think. In addition, *Bloomberg Businessweek* provides a variety of statistics about the state of the economy, including production indexes, stock prices, mutual funds, and interest rates.

Bloomberg Businessweek also uses statistics and statistical information in managing its own business. For example, an annual survey of subscribers helps the company learn about subscriber demographics, reading habits, likely purchases, lifestyles, and so on. *Bloomberg Businessweek* managers use statistical summaries from the survey to provide better services to subscribers and advertisers. One recent North American subscriber survey indicated that 90% of *Bloomberg Businessweek* subscribers use a personal computer at home and that 64% of *Bloomberg Businessweek* subscribers are involved with computer purchases at work. Such statistics alert *Bloomberg Businessweek* managers to subscriber interest in articles about new developments in computers. The results of the subscriber survey are also made available to potential advertisers. The high percentage of subscribers using personal computers at home and the high percentage of subscribers involved with computer purchases at work would be an incentive for a computer manufacturer to consider advertising in *Bloomberg Businessweek.*

In this chapter, we discuss the types of data available for statistical analysis and describe how the data are obtained. We introduce descriptive statistics and statistical inference as ways of converting data into meaningful and easily interpreted statistical information.

AP Photos/Weng lei - Imaginechina

Bloomberg Businessweek uses statistical facts and summaries in many of its articles.

*The authors are indebted to Charlene Trentham, Research Manager, for providing this Statistics in Practice.

Frequently, we see the following types of statements in newspapers and magazines:

- Uber Technologies Inc. is turning to the leveraged-loan market for the first time to raise as much as $2 billion, a sign of the popular ride-sharing network's hunger for cash as it expands around the world (*The Wall Street Journal*, June 14, 2016).
- Against the U.S. dollar, the euro has lost nearly 30% of its value in the last year; the Australian dollar lost almost 20% (*The Economist*, April 25th–May 1st, 2015).

- VW Group's U.S. sales continue to slide, with total sales off by 13% from last January, to 36,930 vehicles (*Panorama*, March 2014).
- A poll of 1320 corporate recruiters indicated that 68% of the recruiters ranked communication skills as one of the top five most important skills for new hires (*Bloomberg Businessweek* April 13–April 19, 2015).
- Green Mountain sold 18 billion coffee pods in two years (*Harvard Business Review*, January-February, 2016).
- Most homeowners spend between about $10,000 and roughly $27,000 converting a basement, depending on the size of the space, according to estimates from HomeAdvisor, a website that connects homeowners with prescreened service professionals (*Consumer Reports*, February 9, 2016).
- A full 88% of consumers say they buy private label, primarily because of price, according to Market Track (*USA Today*, May 17, 2016).

The numerical facts in the preceding statements—$2 billion, 30%, 20%, 13%, 36,930, 1320, 68%, 18 billion, $10,000, $27,000 and 88%—are called **statistics**. In this usage, the term statistics refers to numerical facts such as averages, medians, percentages, and maximums that help us understand a variety of business and economic situations. However, as you will see, the field, or subject, of statistics involves much more than numerical facts. In a broader sense, statistics is the art and science of collecting, analyzing, presenting, and interpreting data. Particularly in business and economics, the information provided by collecting, analyzing, presenting, and interpreting data gives managers and decision makers a better understanding of the business and economic environment and thus enables them to make more informed and better decisions. In this text, we emphasize the use of statistics for business and economic decision making.

Chapter 1 begins with some illustrations of the applications of statistics in business and economics. In Section 1.2 we define the term *data* and introduce the concept of a data set. This section also introduces key terms such as *variables* and *observations,* discusses the difference between quantitative and categorical data, and illustrates the uses of cross-sectional and time series data. Section 1.3 discusses how data can be obtained from existing sources or through survey and experimental studies designed to obtain new data. The important role that the Internet now plays in obtaining data is also highlighted. The uses of data in developing descriptive statistics and in making statistical inferences are described in Sections 1.4 and 1.5. The last four sections of Chapter 1 provide the role of the computer in statistical analysis, an introduction to business analytics and the role statistics plays in it, an introduction to big data and data mining, and a discussion of ethical guidelines for statistical practice.

1.1 Applications in Business and Economics

In today's global business and economic environment, anyone can access vast amounts of statistical information. The most successful managers and decision makers understand the information and know how to use it effectively. In this section, we provide examples that illustrate some of the uses of statistics in business and economics.

Accounting

Public accounting firms use statistical sampling procedures when conducting audits for their clients. For instance, suppose an accounting firm wants to determine whether the amount of accounts receivable shown on a client's balance sheet fairly represents the actual amount of accounts receivable. Usually the large number of individual accounts receivable makes reviewing and validating every account too time-consuming and expensive. As common practice in such situations, the audit staff selects a subset of the accounts called a sample. After reviewing the accuracy of the sampled accounts, the auditors draw a conclusion as to whether the accounts receivable amount shown on the client's balance sheet is acceptable.

Finance

Financial analysts use a variety of statistical information to guide their investment recommendations. In the case of stocks, analysts review financial data such as price/earnings ratios and dividend yields. By comparing the information for an individual stock with information about the stock market averages, an analyst can begin to draw a conclusion as to whether the stock is a good investment. For example, *The Wall Street Journal* (February 27, 2016) reported that the average dividend yield for the S&P 500 companies was 2.3%. Microsoft showed a dividend yield of 2.61%. In this case, the statistical information on dividend yield indicates a higher dividend yield for Microsoft than the average dividend yield for the S&P 500 companies. This and other information about Microsoft would help the analyst make an informed buy, sell, or hold recommendation for Microsoft stock.

Marketing

Electronic scanners at retail checkout counters collect data for a variety of marketing research applications. For example, data suppliers such as ACNielsen and Information Resources, Inc. purchase point-of-sale scanner data from grocery stores, process the data, and then sell statistical summaries of the data to manufacturers. Manufacturers spend hundreds of thousands of dollars per product category to obtain this type of scanner data. Manufacturers also purchase data and statistical summaries on promotional activities such as special pricing and the use of in-store displays. Brand managers can review the scanner statistics and the promotional activity statistics to gain a better understanding of the relationship between promotional activities and sales. Such analyses often prove helpful in establishing future marketing strategies for the various products.

Production

Today's emphasis on quality makes quality control an important application of statistics in production. A variety of statistical quality control charts are used to monitor the output of a production process. In particular, an x-bar chart can be used to monitor the average output. Suppose, for example, that a machine fills containers with 12 ounces of a soft drink. Periodically, a production worker selects a sample of containers and computes the average number of ounces in the sample. This average, or x-bar value, is plotted on an x-bar chart. A plotted value above the chart's upper control limit indicates overfilling, and a plotted value below the chart's lower control limit indicates underfilling. The process is termed "in control" and allowed to continue as long as the plotted x-bar values fall between the chart's upper and lower control limits. Properly interpreted, an x-bar chart can help determine when adjustments are necessary to correct a production process.

Economics

Economists frequently provide forecasts about the future of the economy or some aspect of it. They use a variety of statistical information in making such forecasts. For instance, in forecasting inflation rates, economists use statistical information on such indicators as the Producer Price Index, the unemployment rate, and manufacturing capacity utilization. Often these statistical indicators are entered into computerized forecasting models that predict inflation rates.

Information Systems

Information systems administrators are responsible for the day-to-day operation of an organization's computer networks. A variety of statistical information helps administrators assess the performance of computer networks, including local area networks (LANs), wide area networks (WANs), network segments, intranets, and other data communication systems. Statistics such as the mean number of users on the system, the proportion of time any component of the system is down, and the proportion of bandwidth utilized at various times of the day are examples of statistical information that help the system administrator better understand and manage the computer network.

Applications of statistics such as those described in this section are an integral part of this text. Such examples provide an overview of the breadth of statistical applications. To supplement these examples, practitioners in the fields of business and economics provided chapter-opening Statistics in Practice articles that introduce the material covered in each chapter. The Statistics in Practice applications show the importance of statistics in a wide variety of business and economic situations.

1.2 Data

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation. All the data collected in a particular study are referred to as the **data set** for the study. Table 1.1 shows a data set containing information for 60 nations that participate in the World Trade Organization (WTO). The WTO encourages the free flow of international trade and provides a forum for resolving trade disputes.

Elements, Variables, and Observations

Elements are the entities on which data are collected. Each nation listed in Table 1.1 is an element with the nation or element name shown in the first column. With 60 nations, the data set contains 60 elements.

A **variable** is a characteristic of interest for the elements. The data set in Table 1.1 includes the following five variables:

- WTO Status: The nation's membership status in the World Trade Organization; this can be either as a member or an observer.
- Per Capita GDP ($): The total market value ($) of all goods and services produced by the nation divided by the number of people in the nation; this is commonly used to compare economic productivity of the nations.
- Trade Deficit ($1000s): The difference between the total dollar value of the nation's imports and the total dollar value of the nation's exports.
- Fitch Rating: The nation's sovereign credit rating as appraised by the Fitch Group[1]; the credit ratings range from a high of AAA to a low of F and can be modified by + or −.
- Fitch Outlook: An indication of the direction the credit rating is likely to move over the upcoming two years; the outlook can be negative, stable, or positive.

Measurements collected on each variable for every element in a study provide the data. The set of measurements obtained for a particular element is called an **observation**. Referring to Table 1.1, we see that the first observation contains the following measurements: Member,

[1] The Fitch Group is one of three nationally recognized statistical rating organizations designated by the U.S. Securities and Exchange Commission. The other two are Standard and Poor's and Moody's investor service.

TABLE 1.1 DATA SET FOR 60 NATIONS IN THE WORLD TRADE ORGANIZATION

DATA *file*

Nations

Data sets such as Nations are available on the companion site for this title.

Nation	WTO Status	Per Capita GDP ($)	Trade Deficit ($1000s)	Fitch Rating	Fitch Outlook
Armenia	Member	5,400	2,673,359	BB−	Stable
Australia	Member	40,800	−33,304,157	AAA	Stable
Austria	Member	41,700	12,796,558	AAA	Stable
Azerbaijan	Observer	5,400	−16,747,320	BBB−	Positive
Bahrain	Member	27,300	3,102,665	BBB	Stable
Belgium	Member	37,600	−14,930,833	AA+	Negative
Brazil	Member	11,600	−29,796,166	BBB	Stable
Bulgaria	Member	13,500	4,049,237	BBB−	Positive
Canada	Member	40,300	−1,611,380	AAA	Stable
Cape Verde	Member	4,000	874,459	B+	Stable
Chile	Member	16,100	−14,558,218	A+	Stable
China	Member	8,400	−156,705,311	A+	Stable
Colombia	Member	10,100	−1,561,199	BBB−	Stable
Costa Rica	Member	11,500	5,807,509	BB+	Stable
Croatia	Member	18,300	8,108,103	BBB−	Negative
Cyprus	Member	29,100	6,623,337	BBB	Negative
Czech Republic	Member	25,900	−10,749,467	A+	Positive
Denmark	Member	40,200	−15,057,343	AAA	Stable
Ecuador	Member	8,300	1,993,819	B−	Stable
Egypt	Member	6,500	28,486,933	BB	Negative
El Salvador	Member	7,600	5,019,363	BB	Stable
Estonia	Member	20,200	802,234	A+	Stable
France	Member	35,000	118,841,542	AAA	Stable
Georgia	Member	5,400	4,398,153	B+	Positive
Germany	Member	37,900	−213,367,685	AAA	Stable
Hungary	Member	19,600	−9,421,301	BBB−	Negative
Iceland	Member	38,000	−504,939	BB+	Stable
Ireland	Member	39,500	−59,093,323	BBB+	Negative
Israel	Member	31,000	6,722,291	A	Stable
Italy	Member	30,100	33,568,668	A+	Negative
Japan	Member	34,300	31,675,424	AA	Negative
Kazakhstan	Observer	13,000	−33,220,437	BBB	Positive
Kenya	Member	1,700	9,174,198	B+	Stable
Latvia	Member	15,400	2,448,053	BBB−	Positive
Lebanon	Observer	15,600	13,715,550	B	Stable
Lithuania	Member	18,700	3,359,641	BBB	Positive
Malaysia	Member	15,600	−39,420,064	A−	Stable
Mexico	Member	15,100	1,288,112	BBB	Stable
Peru	Member	10,000	−7,888,993	BBB	Stable
Philippines	Member	4,100	15,667,209	BB+	Stable
Poland	Member	20,100	19,552,976	A−	Stable
Portugal	Member	23,200	21,060,508	BBB−	Negative
South Korea	Member	31,700	−37,509,141	A+	Stable
Romania	Member	12,300	13,323,709	BBB−	Stable
Russia	Observer	16,700	−151,400,000	BBB	Positive
Rwanda	Member	1,300	939,222	B	Stable
Serbia	Observer	10,700	8,275,693	BB−	Stable
Seychelles	Observer	24,700	666,026	B	Stable
Singapore	Member	59,900	−27,110,421	AAA	Stable
Slovakia	Member	23,400	−2,110,626	A+	Stable

Slovenia	Member	29,100	2,310,617	AA−	Negative
South Africa	Member	11,000	3,321,801	BBB+	Stable
Sweden	Member	40,600	−10,903,251	AAA	Stable
Switzerland	Member	43,400	−27,197,873	AAA	Stable
Thailand	Member	9,700	2,049,669	BBB	Stable
Turkey	Member	14,600	71,612,947	BB+	Positive
UK	Member	35,900	162,316,831	AAA	Negative
Uruguay	Member	15,400	2,662,628	BB	Positive
USA	Member	48,100	784,438,559	AAA	Stable
Zambia	Member	1,600	−1,805,198	B+	Stable

5400, 2,673,359, BB−, and Stable. The second observation contains the following measurements: Member, 40,800, −33,304,157, AAA, Stable, and so on. A data set with 60 elements contains 60 observations.

Scales of Measurement

Data collection requires one of the following scales of measurement: nominal, ordinal, interval, or ratio. The scale of measurement determines the amount of information contained in the data and indicates the most appropriate data summarization and statistical analyses.

When the data for a variable consist of labels or names used to identify an attribute of the element, the scale of measurement is considered a **nominal scale**. For example, referring to the data in Table 1.1, the scale of measurement for the WTO Status variable is nominal because the data "member" and "observer" are labels used to identify the status category for the nation. In cases where the scale of measurement is nominal, a numerical code as well as a nonnumerical label may be used. For example, to facilitate data collection and to prepare the data for entry into a computer database, we might use a numerical code for the WTO Status variable by letting 1 denote a member nation in the World Trade Organization and 2 denote an observer nation. The scale of measurement is nominal even though the data appear as numerical values.

The scale of measurement for a variable is considered an **ordinal scale** if the data exhibit the properties of nominal data and in addition, the order or rank of the data is meaningful. For example, referring to the data in Table 1.1, the scale of measurement for the Fitch Rating is ordinal because the rating labels which range from AAA to F can be rank ordered from best credit rating AAA to poorest credit rating F. The rating letters provide the labels similar to nominal data, but in addition, the data can also be ranked or ordered based on the credit rating, which makes the measurement scale ordinal. Ordinal data can also be recorded by a numerical code, for example, your class rank in school.

The scale of measurement for a variable is an **interval scale** if the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numeric. College admission SAT scores are an example of interval-scaled data. For example, three students with SAT math scores of 620, 550, and 470 can be ranked or ordered in terms of best performance to poorest performance in math. In addition, the differences between the scores are meaningful. For instance, student 1 scored $620 - 550 = 70$ points more than student 2, while student 2 scored $550 - 470 = 80$ points more than student 3.

The scale of measurement for a variable is a **ratio scale** if the data have all the properties of interval data and the ratio of two values is meaningful. Variables such as distance, height, weight, and time use the ratio scale of measurement. This scale requires that a zero value be included to indicate that nothing exists for the variable at the zero point. For example, consider the cost of an automobile. A zero value for the cost would indicate that the automobile has no cost and is free. In addition, if we compare the cost of $30,000 for one automobile to the cost of $15,000 for a second automobile, the ratio property shows that the first automobile is $30,000/$15,000 = 2 times, or twice, the cost of the second automobile.

Categorical and Quantitative Data

Data can be classified as either categorical or quantitative. Data that can be grouped by specific categories are referred to as **categorical data**. Categorical data use either the nominal or ordinal scale of measurement. Data that use numeric values to indicate how much or how many are referred to as **quantitative data**. Quantitative data are obtained using either the interval or ratio scale of measurement.

The statistical method appropriate for summarizing data depends upon whether the data are categorical or quantitative.

A **categorical variable** is a variable with categorical data, and a **quantitative variable** is a variable with quantitative data. The statistical analysis appropriate for a particular variable depends upon whether the variable is categorical or quantitative. If the variable is categorical, the statistical analysis is limited. We can summarize categorical data by counting the number of observations in each category or by computing the proportion of the observations in each category. However, even when the categorical data are identified by a numerical code, arithmetic operations such as addition, subtraction, multiplication, and division do not provide meaningful results. Section 2.1 discusses ways of summarizing categorical data.

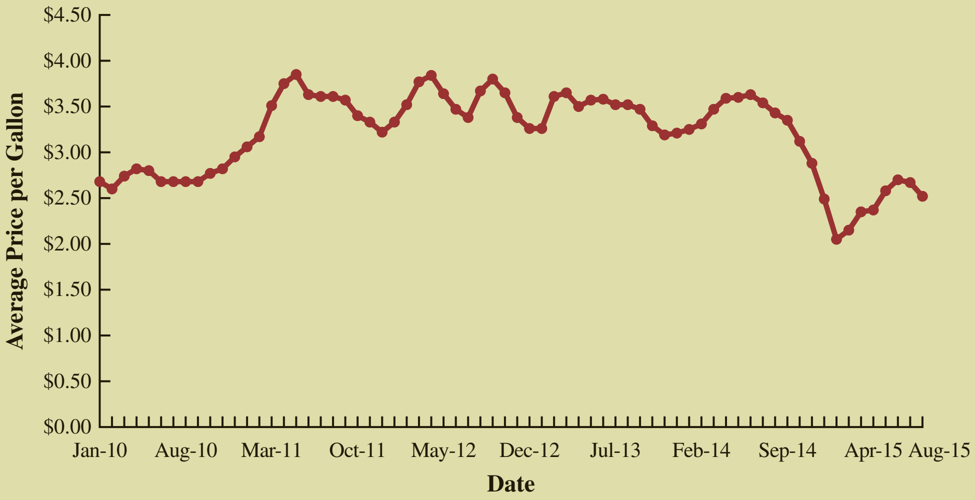
Arithmetic operations provide meaningful results for quantitative variables. For example, quantitative data may be added and then divided by the number of observations to compute the average value. This average is usually meaningful and easily interpreted. In general, more alternatives for statistical analysis are possible when data are quantitative. Section 2.2 and Chapter 3 provide ways of summarizing quantitative data.

Cross-Sectional and Time Series Data

For purposes of statistical analysis, distinguishing between cross-sectional data and time series data is important. **Cross-sectional data** are data collected at the same or approximately the same point in time. The data in Table 1.1 are cross-sectional because they describe the five variables for the 60 World Trade Organization nations at the same point in time. **Time series data** are data collected over several time periods. For example, the time series in Figure 1.1 shows the U.S. average price per gallon of conventional regular gasoline between 2010 and 2015. Note that gasoline prices peaked in May 2011. Between June 2014 and January 2015, the average price per gallon dropped dramatically. In August 2015, the average price per gallon was $2.52.

Graphs of time series data are frequently found in business and economic publications. Such graphs help analysts understand what happened in the past, identify any trends over time, and project future values for the time series. The graphs of time series data can take on a variety of forms, as shown in Figure 1.2. With a little study, these graphs are usually easy to understand and interpret. For example, Panel (A) in Figure 1.2 is a graph that shows the Dow Jones Industrial Average Index from 2005 to 2015. In September 2005, the popular stock market index was near 10,400. Over the next two years the index rose to almost 14,000 in October 2007. However, notice the sharp decline in the time series after the high in 2007. By March 2009, poor economic conditions had caused the Dow Jones Industrial Average Index to return to the 7000 level. This was a scary and discouraging period for investors. However, by late 2009, the index was showing a recovery by reaching 10,000 and rising to a high of over 18,000 in May 2015. By October 2015, the index had dropped substantially to just under 16,300.

FIGURE 1.1 U.S. AVERAGE PRICE PER GALLON FOR CONVENTIONAL REGULAR GASOLINE



Source: Energy Information Administration, U.S. Department of Energy, September 2015.

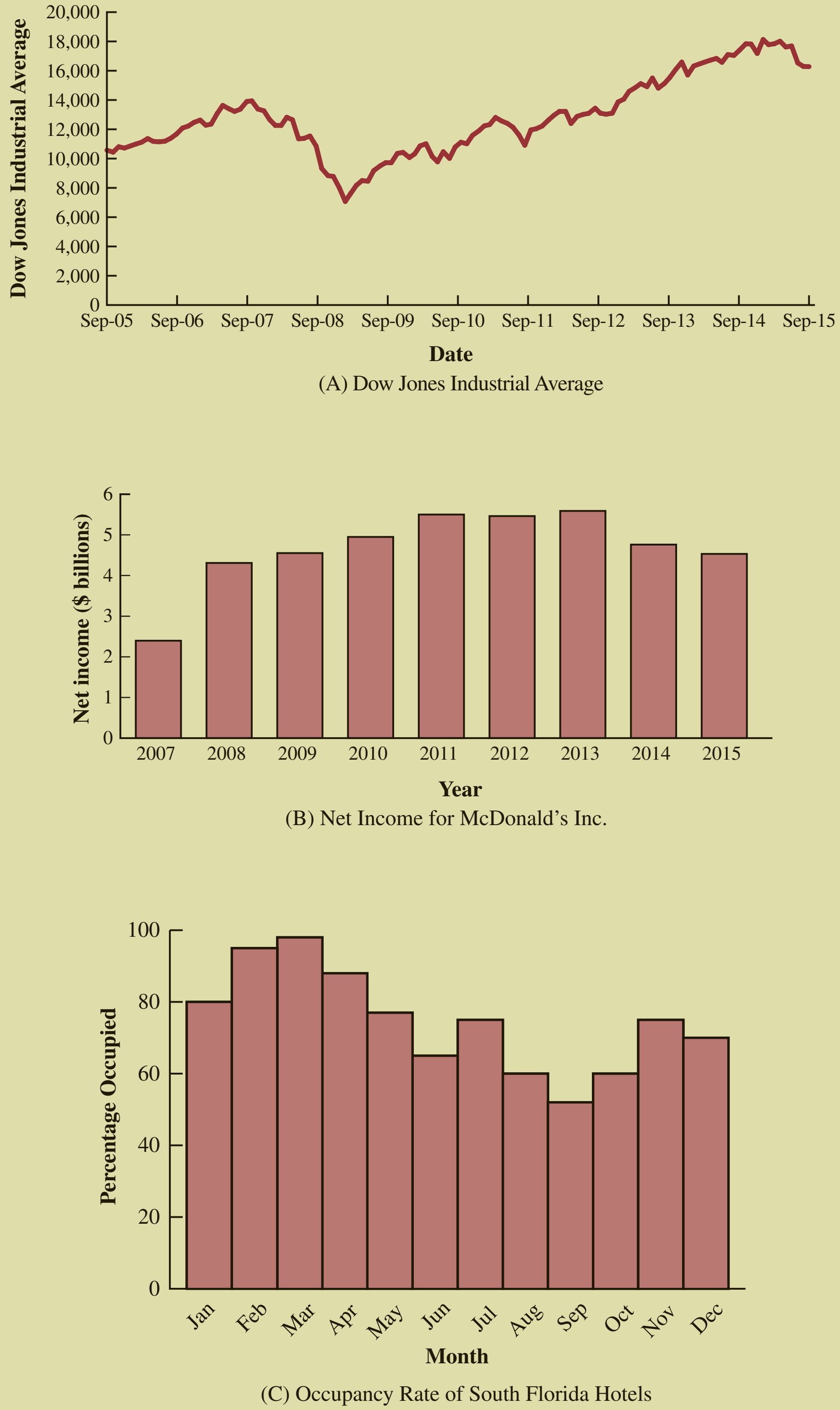
The graph in Panel (B) shows the net income of McDonald's Inc. from 2007 to 2015. The declining economic conditions in 2008 and 2009 were actually beneficial to McDonald's as the company's net income rose to all-time highs. The growth in McDonald's net income showed that the company was thriving during the economic downturn as people were cutting back on the more expensive sit-down restaurants and seeking less expensive alternatives offered by McDonald's. McDonald's net income continued to new all-time highs in 2010 and 2011, remained at about 5.5 billion from 2011 to 2013, decreased substantially in 2014, and dropped again in 2015. Analysts suspect that the drop in net income was due to loss of customers to newer competition such as Chipotle.

Panel (C) shows the time series for the occupancy rate of hotels in South Florida over a one-year period. The highest occupancy rates, 95% and 98%, occur during the months of February and March when the climate of South Florida is attractive to tourists. In fact, January to April of each year is typically the high-occupancy season for South Florida hotels. On the other hand, note the low occupancy rates during the months of August to October, with the lowest occupancy rate of 50% occurring in September. High temperatures and the hurricane season are the primary reasons for the drop in hotel occupancy during this period.

NOTES AND COMMENTS

1. An observation is the set of measurements obtained for each element in a data set. Hence, the number of observations is always the same as the number of elements. The number of measurements obtained for each element equals the number of variables. Hence, the total number of data items can be determined by multiplying the number of observations by the number of variables.
2. Quantitative data may be discrete or continuous. Quantitative data that measure how many (e.g., number of calls received in 5 minutes) are discrete. Quantitative data that measure how much (e.g., weight or time) are continuous because no separation occurs between the possible data values.

FIGURE 1.2 A VARIETY OF GRAPHS OF TIME SERIES DATA



(A) Dow Jones Industrial Average

(B) Net Income for McDonald's Inc.

(C) Occupancy Rate of South Florida Hotels

1.3 Data Sources

Data can be obtained from existing sources, by conducting an observational study, or by conducting an experiment.

Existing Sources

In some cases, data needed for a particular application already exist. Companies maintain a variety of databases about their employees, customers, and business operations. Data on employee salaries, ages, and years of experience can usually be obtained from internal personnel records. Other internal records contain data on sales, advertising expenditures, distribution costs, inventory levels, and production quantities. Most companies also maintain detailed data about their customers. Table 1.2 shows some of the data commonly available from internal company records.

Organizations that specialize in collecting and maintaining data make available substantial amounts of business and economic data. Companies access these external data sources through leasing arrangements or by purchase. Dun & Bradstreet, Bloomberg, and Dow Jones & Company are three firms that provide extensive business database services to clients. ACNielsen and Information Resources, Inc. built successful businesses collecting and processing data that they sell to advertisers and product manufacturers.

Data are also available from a variety of industry associations and special interest organizations. The Travel Industry Association of America maintains travel-related information such as the number of tourists and travel expenditures by states. Such data would be of interest to firms and individuals in the travel industry. The Graduate Management Admission Council maintains data on test scores, student characteristics, and graduate management education programs. Most of the data from these types of sources are available to qualified users at a modest cost.

The Internet is an important source of data and statistical information. Almost all companies maintain websites that provide general information about the company as well as data on sales, number of employees, number of products, product prices, and product specifications. In addition, a number of companies now specialize in making information available over the Internet. As a result, one can obtain access to stock quotes, meal prices at restaurants, salary data, and an almost infinite variety of information.

Government agencies are another important source of existing data. For instance, the U.S. Department of Labor maintains considerable data on employment rates, wage

TABLE 1.2 EXAMPLES OF DATA AVAILABLE FROM INTERNAL COMPANY RECORDS

Source	Some of the Data Typically Available
Employee records	Name, address, social security number, salary, number of vacation days, number of sick days, and bonus
Production records	Part or product number, quantity produced, direct labor cost, and materials cost
Inventory records	Part or product number, number of units on hand, reorder level, economic order quantity, and discount schedule
Sales records	Product number, sales volume, sales volume by region, and sales volume by customer type
Credit records	Customer name, address, phone number, credit limit, and accounts receivable balance
Customer profile	Age, gender, income level, household size, address, and preferences

TABLE 1.3 EXAMPLES OF DATA AVAILABLE FROM SELECTED GOVERNMENT AGENCIES

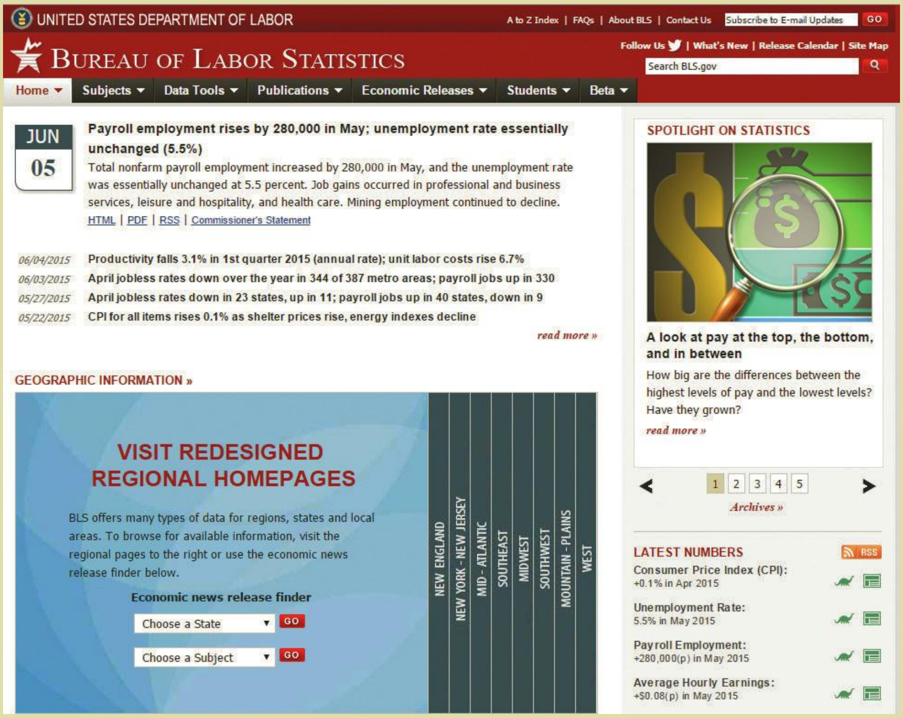
Government Agency	Some of the Data Available
Census Bureau	Population data, number of households, and household income
Federal Reserve Board	Data on the money supply, installment credit, exchange rates, and discount rates
Office of Management and Budget	Data on revenue, expenditures, and debt of the federal government
Department of Commerce	Data on business activity, value of shipments by industry, level of profits by industry, and growing and declining industries
Bureau of Labor Statistics	Consumer spending, hourly earnings, unemployment rate, safety records, and international statistics

rates, size of the labor force, and union membership. Table 1.3 lists selected governmental agencies and some of the data they provide. Most government agencies that collect and process data also make the results available through a website. Figure 1.3 shows the homepage for the U.S. Bureau of Labor Statistics website.

Observational Study

In an *observational study* we simply observe what is happening in a particular situation, record data on one or more variables of interest, and conduct a statistical analysis of the resulting data. For example, researchers might observe a randomly selected group of

FIGURE 1.3 U.S. BUREAU OF LABOR STATISTICS HOMEPAGE



Payroll employment rises by 280,000 in May; unemployment rate essentially unchanged (5.5%)

Total nonfarm payroll employment increased by 280,000 in May, and the unemployment rate was essentially unchanged at 5.5 percent. Job gains occurred in professional and business services, leisure and hospitality, and health care. Mining employment continued to decline.

HTML | PDF | RSS | Commissioner's Statement

06/04/2015	Productivity falls 3.1% in 1st quarter 2015 (annual rate); unit labor costs rise 6.7%
06/03/2015	April jobless rates down over the year in 344 of 387 metro areas; payroll jobs up in 330
05/27/2015	April jobless rates down in 23 states, up in 11; payroll jobs up in 40 states, down in 9
05/22/2015	CPI for all items rises 0.1% as shelter prices rise, energy indexes decline

read more »

A look at pay at the top, the bottom, and in between

How big are the differences between the highest levels of pay and the lowest levels? Have they grown?

read more »

VISIT REDESIGNED REGIONAL HOMEPAGES

BLS offers many types of data for regions, states and local areas. To browse for available information, visit the regional pages to the right or use the economic news release finder below.

Economic news release finder

Consumer Price Index (CPI):
+0.1% in Apr 2015

Unemployment Rate:
5.5% in May 2015

Payroll Employment:
+280,000(p) in May 2015

Average Hourly Earnings:
+$0.08(p) in May 2015

Courtesy of U.S. Bureau of Labor Statistics

Studies of smokers and nonsmokers are observational studies because researchers do not determine or control who will smoke and who will not smoke.

customers that enter a Walmart supercenter to collect data on variables such as the length of time the customer spends shopping, the gender of the customer, the amount spent, and so on. Statistical analysis of the data may help management determine how factors such as the length of time shopping and the gender of the customer affect the amount spent.

As another example of an observational study, suppose that researchers were interested in investigating the relationship between the gender of the CEO for a *Fortune* 500 company and the performance of the company as measured by the return on equity (ROE). To obtain data, the researchers selected a sample of companies and recorded the gender of the CEO and the ROE for each company. Statistical analysis of the data can help determine the relationship between performance of the company and the gender of the CEO. This example is an observational study because the researchers had no control over the gender of the CEO or the ROE at each of the companies that were sampled.

Surveys and public opinion polls are two other examples of commonly used observational studies. The data provided by these types of studies simply enable us to observe opinions of the respondents. For example, the New York State legislature commissioned a telephone survey in which residents were asked if they would support or oppose an increase in the state gasoline tax in order to provide funding for bridge and highway repairs. Statistical analysis of the survey results will assist the state legislature in determining if it should introduce a bill to increase gasoline taxes.

Experiment

The largest experimental statistical study ever conducted is believed to be the 1954 Public Health Service experiment for the Salk polio vaccine. Nearly 2 million children in grades 1, 2, and 3 were selected from throughout the United States.

The key difference between an observational study and an experiment is that an experiment is conducted under controlled conditions. As a result, the data obtained from a well-designed experiment can often provide more information as compared to the data obtained from existing sources or by conducting an observational study. For example, suppose a pharmaceutical company would like to learn about how a new drug it has developed affects blood pressure. To obtain data about how the new drug affects blood pressure, researchers selected a sample of individuals. Different groups of individuals are given different dosage levels of the new drug, and before and after data on blood pressure are collected for each group. Statistical analysis of the data can help determine how the new drug affects blood pressure.

The types of experiments we deal with in statistics often begin with the identification of a particular variable of interest. Then one or more other variables are identified and controlled so that data can be obtained about how the other variables influence the primary variable of interest. In Chapter 13 we discuss statistical methods appropriate for analyzing the data from an experiment.

Time and Cost Issues

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

Data Acquisition Errors

Managers should always be aware of the possibility of data errors in statistical studies. Using erroneous data can be worse than not using any data at all. An error in data acquisition occurs whenever the data value obtained is not equal to the true or actual value that would be obtained with a correct procedure. Such errors can occur in a number of ways. For example, an interviewer might make a recording error, such as a transposition in writing the age of a 24-year-old person as 42, or the person answering an interview question might misinterpret the question and provide an incorrect response.

Experienced data analysts take great care in collecting and recording data to ensure that errors are not made. Special procedures can be used to check for internal consistency of the data. For instance, such procedures would indicate that the analyst should review the accuracy of data for a respondent shown to be 22 years of age but reporting 20 years of work experience. Data analysts also review data with unusually large and small values, called outliers, which are candidates for possible data errors. In Chapter 3 we present some of the methods statisticians use to identify outliers.

Errors often occur during data acquisition. Blindly using any data that happen to be available or using data that were acquired with little care can result in misleading information and bad decisions. Thus, taking steps to acquire accurate data can help ensure reliable and valuable decision-making information.

1.4 Descriptive Statistics

Most of the statistical information in newspapers, magazines, company reports, and other publications consists of data that are summarized and presented in a form that is easy for the reader to understand. Such summaries of data, which may be tabular, graphical, or numerical, are referred to as **descriptive statistics**.

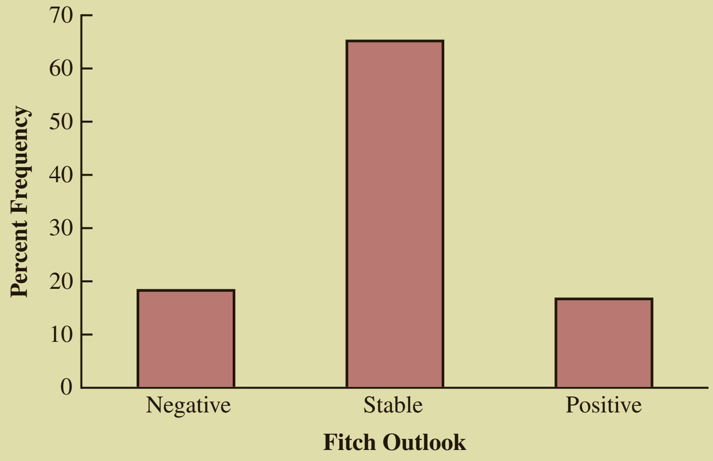
Refer to the data set in Table 1.1 showing data for 60 nations that participate in the World Trade Organization. Methods of descriptive statistics can be used to summarize these data. For example, consider the variable Fitch Outlook, which indicates the direction the nation's credit rating is likely to move over the next two years. The Fitch Outlook is recorded as being negative, stable, or positive. A tabular summary of the data showing the number of nations with each of the Fitch Outlook ratings is shown in Table 1.4. A graphical summary of the same data, called a bar chart, is shown in Figure 1.4. These types of summaries make the data easier to interpret. Referring to Table 1.4 and Figure 1.4, we can see that the majority of Fitch Outlook credit ratings are stable, with 65% of the nations having this rating. Negative and positive outlook credit ratings are similar, with slightly more nations having a negative outlook (18.3%) than a positive outlook (16.7%).

A graphical summary of the data for quantitative variable Per Capita GDP in Table 1.1, called a histogram, is provided in Figure 1.5. Using the histogram, it is easy to see that Per Capita GDP for the 60 nations ranges from $0 to $60,000, with the highest concentration between $10,000 and $20,000. Only one nation had a Per Capita GDP exceeding $50,000.

TABLE 1.4 FREQUENCIES AND PERCENT FREQUENCIES FOR THE FITCH CREDIT RATING OUTLOOK OF 60 NATIONS

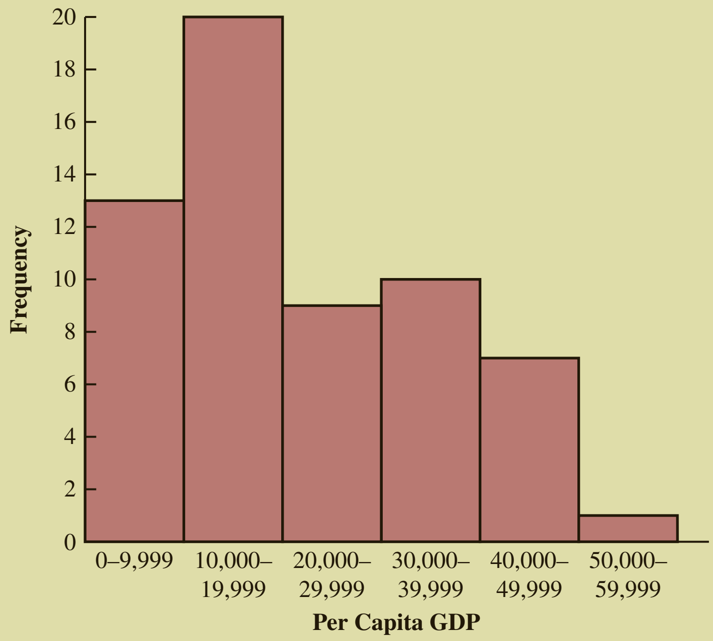
Fitch Outlook	Frequency	Percent Frequency (%)
Positive	10	16.7
Stable	39	65.0
Negative	11	18.3

FIGURE 1.4 BAR CHART FOR THE FITCH CREDIT RATING OUTLOOK FOR 60 NATIONS



In addition to tabular and graphical displays, numerical descriptive statistics are used to summarize data. The most common numerical measure is the average, or mean. Using the data on Per Capita GDP for the 60 nations in Table 1.1, we can compute the average by adding Per Capita GDP for all 60 nations and dividing the total by 60. Doing so provides an average Per Capita GDP of $21,387. This average provides a measure of the central tendency, or central location of the data.

FIGURE 1.5 HISTOGRAM OF PER CAPITA GDP FOR 60 NATIONS



There is a great deal of interest in effective methods for developing and presenting descriptive statistics. Chapters 2 and 3 devote attention to the tabular, graphical, and numerical methods of descriptive statistics.

1.5 Statistical Inference

Many situations require information about a large group of elements (individuals, companies, voters, households, products, customers, and so on). But, because of time, cost, and other considerations, data can be collected from only a small portion of the group. The larger group of elements in a particular study is called the **population**, and the smaller group is called the **sample**. Formally, we use the following definitions.

POPULATION

A population is the set of all elements of interest in a particular study.

SAMPLE

A sample is a subset of the population.

The U.S. government conducts a census every 10 years. Market research firms conduct sample surveys every day.

The process of conducting a survey to collect data for the entire population is called a **census**. The process of conducting a survey to collect data for a sample is called a **sample survey**. As one of its major contributions, statistics uses data from a sample to make estimates and test hypotheses about the characteristics of a population through a process referred to as **statistical inference**.

As an example of statistical inference, let us consider the study conducted by Norris Electronics. Norris manufactures a high-intensity lightbulb used in a variety of electrical products. In an attempt to increase the useful life of the lightbulb, the product design group developed a new lightbulb filament. In this case, the population is defined as all lightbulbs that could be produced with the new filament. To evaluate the advantages of the new filament, 200 bulbs with the new filament were manufactured and tested. Data collected from this sample showed the number of hours each lightbulb operated before filament burnout. See Table 1.5.

Suppose Norris wants to use the sample data to make an inference about the average hours of useful life for the population of all lightbulbs that could be produced with the new filament. Adding the 200 values in Table 1.5 and dividing the total by 200 provides the sample average lifetime for the lightbulbs: 76 hours. We can use this sample result to estimate that the average lifetime for the lightbulbs in the population is 76 hours. Figure 1.6 provides a graphical summary of the statistical inference process for Norris Electronics.

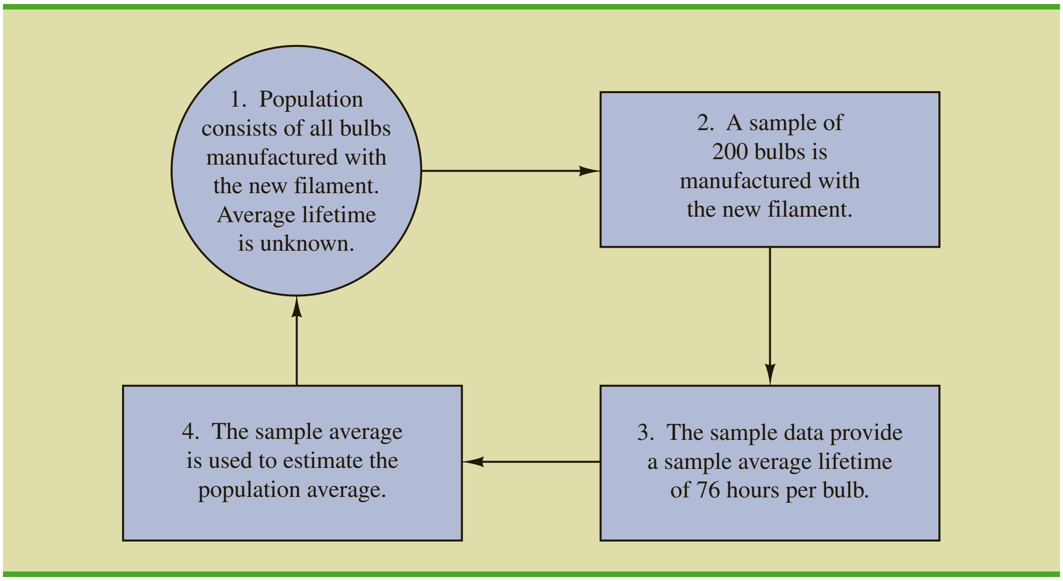
Whenever statisticians use a sample to estimate a population characteristic of interest, they usually provide a statement of the quality, or precision, associated with the estimate. For the Norris example, the statistician might state that the point estimate of the average lifetime for the population of new lightbulbs is 76 hours with a margin of error of ±4 hours. Thus, an interval estimate of the average lifetime for all lightbulbs produced with the new filament is 72 hours to 80 hours. The statistician can also state how confident he or she is that the interval from 72 hours to 80 hours contains the population average.

TABLE 1.5 HOURS UNTIL BURNOUT FOR A SAMPLE OF 200 LIGHTBULBS FOR THE NORRIS ELECTRONICS EXAMPLE

DATA file
Norris

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73

FIGURE 1.6 THE PROCESS OF STATISTICAL INFERENCE FOR THE NORRIS ELECTRONICS EXAMPLE


1. Population consists of all bulbs manufactured with the new filament. Average lifetime is unknown.
2. A sample of 200 bulbs is manufactured with the new filament.
3. The sample data provide a sample average lifetime of 76 hours per bulb.
4. The sample average is used to estimate the population average.

1.6 Statistical Analysis Using Microsoft Excel

Because statistical analysis typically involves working with large amounts of data, computer software is frequently used to conduct the analysis. In this book we show how statistical analysis can be performed using Microsoft Excel.

We want to emphasize that this book is about statistics; it is not a book about spreadsheets. Our focus is on showing the appropriate statistical procedures for collecting, analyzing, presenting, and interpreting data. Because Excel is widely available in business organizations, you can expect to put the knowledge gained here to use in the setting where you currently, or soon will, work. If, in the process of studying this material, you become more proficient with Excel, so much the better.

We begin most sections with an application scenario in which a statistical procedure is useful. After showing what the statistical procedure is and how it is used, we turn to showing how to implement the procedure using Excel. Thus, you should gain an understanding of what the procedure is, the situation in which it is useful, and how to implement it using the capabilities of Excel.

Data Sets and Excel Worksheets

Data sets are organized in Excel worksheets in much the same way as the data set for the 60 nations that participate in the World Trade Organization that appears in Table 1.1 is organized. Figure 1.7 shows an Excel worksheet for that data set. Note that row 1 and column A contain labels. Cells Bl:Fl contain the variable names; cells A2:A61 contain the observation names; and cells B2:F61 contain the data that were collected. A purple fill color is used to highlight the cells that contain the data. Displaying a worksheet with this many rows on a single page of a textbook is not practical. In such cases we will hide selected rows to conserve space. In the Excel worksheet shown in Figure 1.7 we have hidden rows 15 through 54 (observations 14 through 53) to conserve space.

To hide rows 15 through 54 of the Excel worksheet, first select rows 15 through 54. Then, right-click and choose the Hide option. To redisplay rows 15 through 54, just select rows 14 through 55, right-click, and select the Unhide option.

The data are the focus of the statistical analysis. Except for the headings in row 1, each row of the worksheet corresponds to an observation and each column corresponds to a variable. For instance, row 2 of the worksheet contains the data for the first observation, Armenia; row 3 contains the data for the second observation, Australia; row 3 contains the data for the third observation, Austria; and so on. The names in column A provide a convenient way to refer to each of the 60 observations in the study. Note that column B of the worksheet contains the data for the variable WTO Status, column C contains the data for the Per Capita GDP ($), and so on.

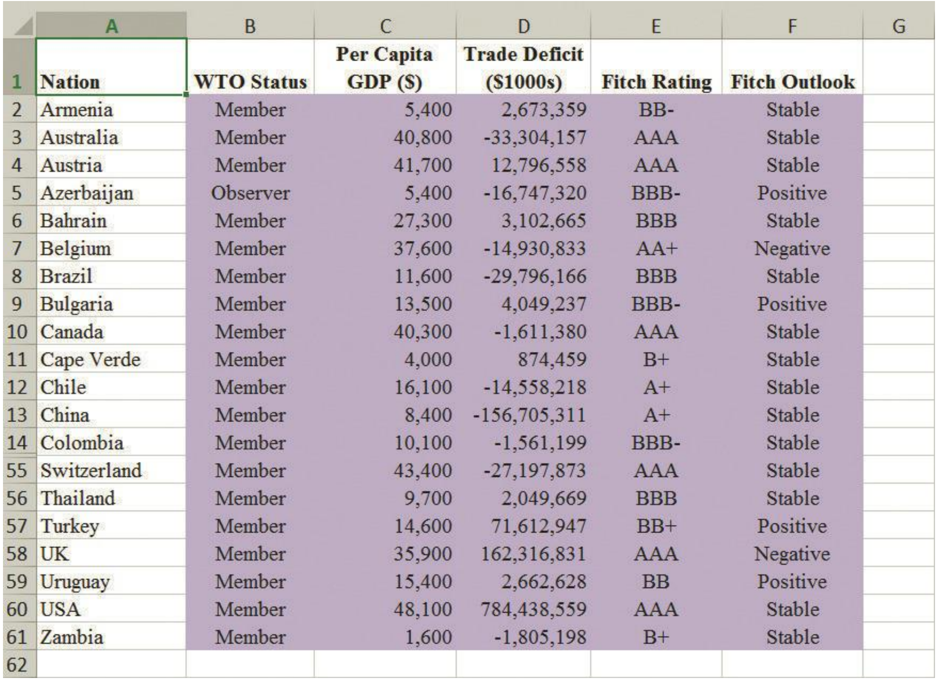
Suppose now that we want to use Excel to analyze the Norris Electronics data shown in Table 1.5. The data in Table 1.5 are organized into 10 columns with 20 data values in each column so that the data would fit nicely on a single page of the text. Even though the table has several columns, it shows data for only one variable (hours until burnout). In statistical worksheets it is customary to put all the data for each variable in a single column. Refer to the Excel worksheet shown in Figure 1.8. To make it easier to identify each observation in the data set, we entered the heading Observation into cell Al and the numbers 1–200 into cells A2:A201. The heading Hours Until Burnout has been entered into cell B1, and the data for the 200 observations have been entered into cells B2:B201. Note that rows 7 through 195 have been hidden to conserve space.

Using Excel for Statistical Analysis

To separate the discussion of a statistical procedure from the discussion of using Excel to implement the procedure, the material that discusses the use of Excel will usually be set apart in sections with headings such as Using Excel to Construct a Bar Chart and a Pie

FIGURE 1.7 EXCEL WORKSHEET FOR THE 60 NATIONS THAT PARTICIPATE IN THE WORLD TRADE ORGANIZATION

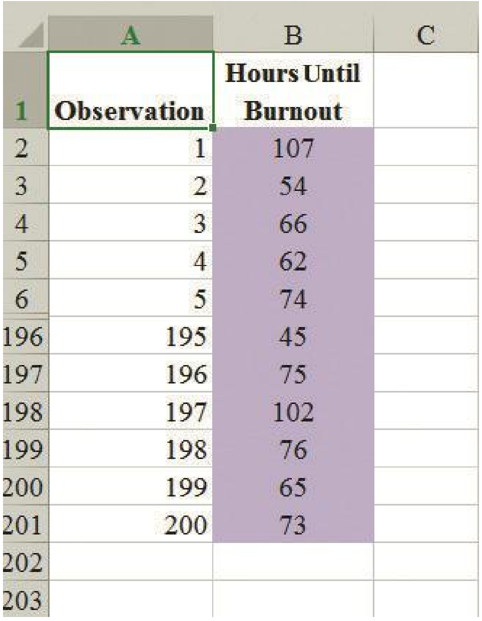
Note: Rows 15–54 are hidden.



	A	B	C	D	E	F	G
1	**Nation**	**WTO Status**	**Per Capita GDP ($)**	**Trade Deficit ($1000s)**	**Fitch Rating**	**Fitch Outlook**	
2	Armenia	Member	5,400	2,673,359	BB-	Stable	
3	Australia	Member	40,800	-33,304,157	AAA	Stable	
4	Austria	Member	41,700	12,796,558	AAA	Stable	
5	Azerbaijan	Observer	5,400	-16,747,320	BBB-	Positive	
6	Bahrain	Member	27,300	3,102,665	BBB	Stable	
7	Belgium	Member	37,600	-14,930,833	AA+	Negative	
8	Brazil	Member	11,600	-29,796,166	BBB	Stable	
9	Bulgaria	Member	13,500	4,049,237	BBB-	Positive	
10	Canada	Member	40,300	-1,611,380	AAA	Stable	
11	Cape Verde	Member	4,000	874,459	B+	Stable	
12	Chile	Member	16,100	-14,558,218	A+	Stable	
13	China	Member	8,400	-156,705,311	A+	Stable	
14	Colombia	Member	10,100	-1,561,199	BBB-	Stable	
55	Switzerland	Member	43,400	-27,197,873	AAA	Stable	
56	Thailand	Member	9,700	2,049,669	BBB	Stable	
57	Turkey	Member	14,600	71,612,947	BB+	Positive	
58	UK	Member	35,900	162,316,831	AAA	Negative	
59	Uruguay	Member	15,400	2,662,628	BB	Positive	
60	USA	Member	48,100	784,438,559	AAA	Stable	
61	Zambia	Member	1,600	-1,805,198	B+	Stable	
62							

FIGURE 1.8 EXCEL WORKSHEET FOR THE NORRIS ELECTRONICS DATA SET

Note: Rows 7–195 are hidden



	A	B	C
1	**Observation**	**Hours Until Burnout**	
2	1	107	
3	2	54	
4	3	66	
5	4	62	
6	5	74	
196	195	45	
197	196	75	
198	197	102	
199	198	76	
200	199	65	
201	200	73	
202			
203			

Chart, Using Excel to Construct a Frequency Distribution, and so on. In using Excel for statistical analysis, four tasks may be needed: Enter/Access Data; Enter Functions and Formulas; Apply Tools; and Editing Options.

Enter/Access Data: Select cell locations for the data and enter the data along with appropriate labels; or open an existing Excel file such as one of the DATAfiles that accompany the text.

Enter Functions and Formulas: Select cell locations, enter Excel functions and formulas, and provide descriptive labels to identify the results.

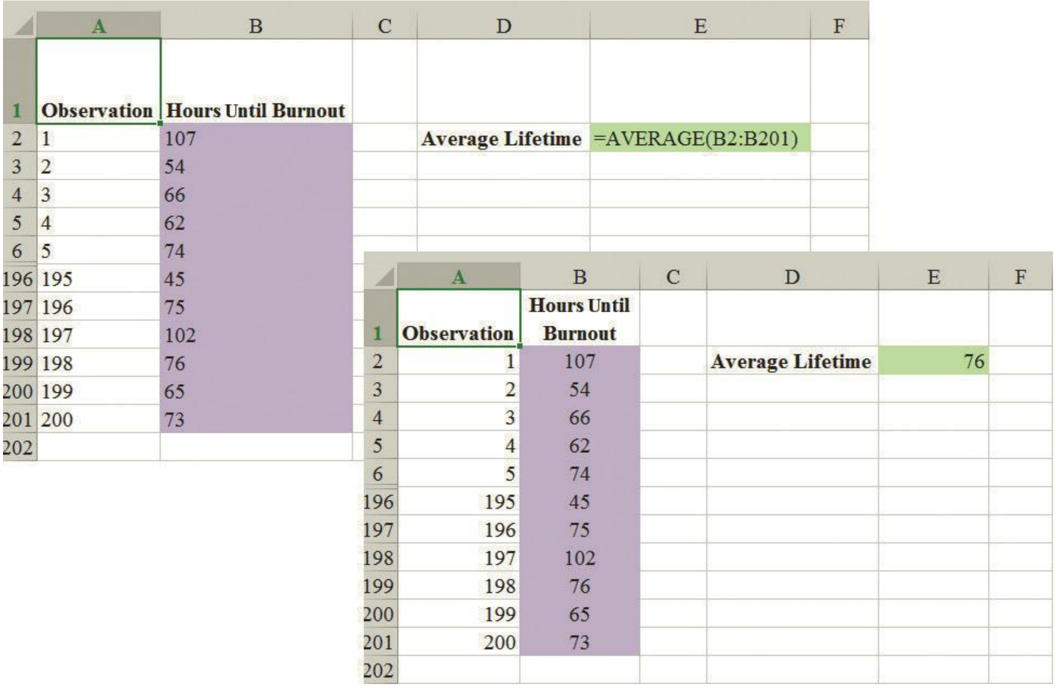
Apply Tools: Use Excel's tools for data analysis and presentation.

Editing Options: Edit the results to better identify the output or to create a different type of presentation. For example, when using Excel's chart tools, we can edit the chart that is created by adding, removing, or changing chart elements such as the title, legend, data labels, and so on.

Our approach will be to describe how these tasks are performed each time we use Excel to implement a statistical procedure. It will always be necessary to enter data or open an existing Excel file. But, depending on the complexity of the statistical analysis, only one of the second or third tasks may be needed.

To illustrate how the discussion of Excel will appear throughout the book, we will show how to use Excel's AVERAGE function to compute the average lifetime for the 200 burnout times in Table 1.5. Refer to Figure 1.9 as we describe the tasks involved. The worksheet shown in the foreground of Figure 1.9 displays the data for the problem and shows the results of the analysis. It is called the *value worksheet*. The worksheet shown in the

FIGURE 1.9 COMPUTING THE AVERAGE LIFETIME OF LIGHTBULBS FOR NORRIS ELECTRONICS USING EXCEL'S AVERAGE FUNCTION



	A	B	C	D	E	F
1	**Observation**	**Hours Until Burnout**				
2	1	107		**Average Lifetime**	=AVERAGE(B2:B201)	
3	2	54				
4	3	66				
5	4	62				
6	5	74				
196	195	45				
197	196	75				
198	197	102				
199	198	76				
200	199	65				
201	200	73				
202						

	A	B	C	D	E	F
1	**Observation**	**Hours Until Burnout**				
2	1	107		**Average Lifetime**	76	
3	2	54				
4	3	66				
5	4	62				
6	5	74				
196	195	45				
197	196	75				
198	197	102				
199	198	76				
200	199	65				
201	200	73				
202						

background displays the Excel formula used to compute the average lifetime and is called the *formula worksheet*. A purple fill color is used to highlight the cells that contain the data in both worksheets. In addition, a green fill color is used to highlight the cells containing the functions and formulas in the formula worksheet and the corresponding results in the value worksheet.

Enter/Access Data: Open the DATAfile named *Norris*. The data are in cells B2:B201 and labels are in column A and cell B1.

Enter Functions and Formulas: Excel's AVERAGE function can be used to compute the mean by entering the following formula into cell E2:

=AVERAGE(B2:201)

Similarly, the formulas =MEDIAN(B2:B201) and =MODE.SNGL(B2:B201) are entered into cells E3 and E4, respectively, to compute the median and the mode.

To identify the result, the label Average Lifetime is entered into cell D2. Note that for this illustration the Apply Tools and Editing Options tasks were not required. The value worksheet shows that the value computed using the AVERAGE function is 76 hours.

1.7 Analytics

Because of the dramatic increase in available data, more cost-effective data storage, faster computer processing, and recognition by managers that data can be extremely valuable for understanding customers and business operations, there has been a dramatic increase in data-driven decision making. The broad range of techniques that may be used to support data-driven decisions comprise what has become known as analytics.

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

Analytics is the scientific process of transforming data into insight for making better decisions. Analytics is used for data-driven or fact-based decision making, which is often seen as more objective than alternative approaches to decision making. The tools of analytics can aid decision making by creating insights from data, improving our ability to more accurately forecast for planning, helping us quantify risk, and yielding better alternatives through analysis.

Analytics can involve a variety of techniques from simple reports to the most advanced optimization techniques (algorithms for finding the best course of action). Analytics is now generally thought to comprise three broad categories of techniques. These categories are descriptive analytics, predictive analytics, and prescriptive analytics.

Descriptive analytics encompasses the set of analytical techniques that describe what has happened in the past. Examples of these types of techniques are data queries, reports, descriptive statistics, data visualization, data dash boards, and basic what-if spreadsheet models.

Predictive analytics consists of analytical techniques that use models constructed from past data to predict the future or to assess the impact of one variable on another. For example, past data on sales of a product may be used to construct a mathematical model that predicts future sales. Such a model can account for factors such as the growth trajectory and seasonality of the product's sales based on past growth and seasonal patterns. Point-of-sale scanner data from retail outlets may be used by a packaged food manufacturer to help estimate the lift in unit sales associated with coupons or sales events. Survey data and past purchase behavior may be used to help predict the market share of a new product. Each of these is an example of predictive analytics. Linear regression, time series analysis, and forecasting models fall into the category of predictive analytics; these techniques are discussed later in this text. Simulation, which is the use of probability and statistical computer models to better understand risk, also falls under the category of predictive analytics.

Prescriptive analytics differs greatly from descriptive or predictive analytics. What distinguishes prescriptive analytics is that prescriptive models yield a best course of action to take. That is, the output of a prescriptive model is a best decision. Hence, **prescriptive analytics** is the set of analytical techniques that yield a course of action. Optimization models, which generate solutions that maximize or minimize some objective subject to a set of constraints, fall into the category of prescriptive models. The airline industry's use of revenue management is an example of a prescriptive model. The airline industry uses past purchasing data as inputs into a model that recommends the pricing strategy across all flights that will maximize revenue for the company.

How does the study of statistics relate to analytics? Most of the techniques in descriptive and predictive analytics come from probability and statistics. These include descriptive statistics, data visualization, probability and probability distributions, sampling, and predictive modeling, including regression analysis and time series forecasting. Each of these techniques is discussed in this text. The increased use of analytics for data-driven decision making makes it more important than ever for analysts and managers to understand statistics and data analysis. Companies are increasingly seeking data savvy managers who know how to use descriptive and predictive models to make data-driven decisions.

At the beginning of this section, we mentioned the increased availability of data as one of the drivers of the interest in analytics. In the next section we discuss this explosion in available data and how it relates to the study of statistics.

1.8 Big Data and Data Mining

With the aid of magnetic card readers, bar code scanners, and point-of-sale terminals, most organizations obtain large amounts of data on a daily basis. And, even for a small local restaurant that uses touch screen monitors to enter orders and handle billing, the amount of data collected can be substantial. For large retail companies, the sheer volume of data collected is hard to conceptualize, and figuring out how to effectively use these data to improve profitability is a challenge. Mass retailers such as Walmart capture data on 20 to 30 million transactions every day, telecommunication companies such as France Telecom and AT&T generate over 300 million call records per day, and Visa processes 6800 payment transactions per second or approximately 600 million transactions per day.

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

Larger and more complex data sets are now often referred to as **big data**. Although there does not seem to be a universally accepted definition of *big data*, many think if it as a set of data that cannot be managed, processed, or analyzed with commonly available software in a reasonable amount of time. Many data analysts define *big data* by referring to the three v's of data: volume, velocity, and variety. *Volume* refers to the amount of available data (the typical unit of measure for data is now a terabyte, which is 10^{12} bytes); *velocity* refers to the speed at which data is collected and processed; and *variety* refers to the different data types.

The term *data warehousing* is used to refer to the process of capturing, storing, and maintaining the data. Computing power and data collection tools have reached the point where it is now feasible to store and retrieve extremely large quantities of data in seconds. Analysis of the data in the warehouse may result in decisions that will lead to new strategies and higher profits for the organization. For example, General Electric (GE) captures a large amount of data from sensors on its aircraft engines each time a plane takes off or lands. Capturing these data allows GE to offer an important service to its customers; GE monitors the engine performance and can alert its customer when service is needed or a problem is likely to occur.

The subject of **data mining** deals with methods for developing useful decision-making information from large databases. Using a combination of procedures from statistics, mathematics, and computer science, analysts "mine the data" in the warehouse to convert it into useful information, hence the name *data mining.* Dr. Kurt Thearling, a leading practitioner in the field, defines data mining as "the automated extraction of predictive information from (large) databases." The two key words in Dr. Thearling's definition are "automated" and "predictive." Data mining systems that are the most effective use automated procedures to extract information from the data using only the most general or even vague queries by the user. And data mining software automates the process of uncovering hidden predictive information that in the past required hands-on analysis.

The major applications of data mining have been made by companies with a strong consumer focus, such as retail businesses, financial organizations, and communication companies. Data mining has been successfully used to help retailers such as Amazon and Barnes & Noble determine one or more related products that customers who have already purchased a specific product are also likely to purchase. Then, when a customer logs on to the company's website and purchases a product, the website uses pop-ups to alert the customer about additional products that the customer is likely to purchase. In another application, data mining may be used to identify customers who are likely to spend more than $20 on a particular shopping trip. These customers may then be identified as the ones to receive special e-mail or regular mail discount offers to encourage them to make their next shopping trip before the discount termination date.

Statistical methods play an important role in data mining, both in terms of discovering relationships in the data and predicting future outcomes. However, a thorough coverage of data mining and the use of statistics in data mining is outside the scope of this text.

Data mining is a technology that relies heavily on statistical methodology such as multiple regression, logistic regression, and correlation. But it takes a creative integration of all these methods and computer science technologies involving artificial intelligence and machine learning to make data mining effective. A substantial investment in time and money is required to implement commercial data mining software packages developed by firms such as Oracle, Teradata, and SAS. The statistical concepts introduced in this text will be helpful in understanding the statistical methodology used by data mining software packages and enable you to better understand the statistical information that is developed.

Because statistical models play an important role in developing predictive models in data mining, many of the concerns that statisticians deal with in developing statistical models are also applicable. For instance, a concern in any statistical study involves the issue of model reliability. Finding a statistical model that works well for a particular sample of data does not necessarily mean that it can be reliably applied to other data. One of the common statistical approaches to evaluating model reliability is to divide the sample data set into two parts: a training data set and a test data set. If the model developed using the training data is able to accurately predict values in the test data, we say that the model is reliable. One advantage that data mining has over classical statistics is that the enormous amount of data available allows the data mining software to partition the data set so that a model developed for the training data set may be tested for reliability on other data. In this sense, the partitioning of the data set allows data mining to develop models and relationships and then quickly observe if they are repeatable and valid with new and different data. On the other hand, a warning for data mining applications is that with so much data available, there is a danger of overfitting the model to the point that misleading associations and cause/effect conclusions appear to exist. Careful interpretation of data mining results and additional testing will help avoid this pitfall.

1.9 Ethical Guidelines for Statistical Practice

Ethical behavior is something we should strive for in all that we do. Ethical issues arise in statistics because of the important role statistics plays in the collection, analysis, presentation, and interpretation of data. In a statistical study, unethical behavior can take a variety of forms including improper sampling, inappropriate analysis of the data, development of misleading graphs, use of inappropriate summary statistics, and/or a biased interpretation of the statistical results.

As you begin to do your own statistical work, we encourage you to be fair, thorough, objective, and neutral as you collect data, conduct analyses, make oral presentations, and present written reports containing information developed. As a consumer of statistics, you should also be aware of the possibility of unethical statistical behavior by others. When you see statistics in newspapers, on television, on the Internet, and so on, it is a good idea to view the information with some skepticism, always being aware of the source as well as the purpose and objectivity of the statistics provided.

The American Statistical Association, the nation's leading professional organization for statistics and statisticians, developed the report "Ethical Guidelines for Statistical Practice"[2] to help statistical practitioners make and communicate ethical decisions and assist students in learning how to perform statistical work responsibly. The report contains 67 guidelines organized into eight topic areas: Professionalism; Responsibilities to Funders, Clients, and Employers; Responsibilities in Publications and Testimony; Responsibilities to Research Subjects; Responsibilities to Research Team Colleagues; Responsibilities to Other Statisticians or Statistical Practitioners; Responsibilities Regarding Allegations of Misconduct; and Responsibilities of Employers Including Organizations, Individuals, Attorneys, or Other Clients Employing Statistical Practitioners.

One of the ethical guidelines in the professionalism area addresses the issue of running multiple tests until a desired result is obtained. Let us consider an example. In Section 1.5 we discussed a statistical study conducted by Norris Electronics involving a sample of 200 high-intensity lightbulbs manufactured with a new filament. The average lifetime for the sample, 76 hours, provided an estimate of the average lifetime for all lightbulbs produced with the new filament. However, consider this. Because Norris selected a sample of bulbs, it is reasonable to assume that another sample would have provided a different average lifetime.

Suppose Norris's management had hoped the sample results would enable them to claim that the average lifetime for the new lightbulbs was 80 hours or more. Suppose further that Norris's management decides to continue the study by manufacturing and testing repeated samples of 200 lightbulbs with the new filament until a sample mean of 80 hours or more is obtained. If the study is repeated enough times, a sample may eventually be obtained—by chance alone—that would provide the desired result and enable Norris to make such a claim. In this case, consumers would be misled into thinking the new product is better than it actually is. Clearly, this type of behavior is unethical and represents a gross misuse of statistics in practice.

Several ethical guidelines in the responsibilities and publications and testimony area deal with issues involving the handling of data. For instance, a statistician must account for

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

all data considered in a study and explain the sample(s) actually used. In the Norris Electronics study the average lifetime for the 200 bulbs in the original sample is 76 hours; this is considerably less than the 80 hours or more that management hoped to obtain. Suppose now that after reviewing the results showing a 76-hour average lifetime, Norris discards all the observations with 70 or fewer hours until burnout, allegedly because these bulbs contain imperfections caused by startup problems in the manufacturing process. After these lightbulbs are discarded, the average lifetime for the remaining lightbulbs in the sample turns out to be 82 hours. Would you be suspicious of Norris's claim that the lifetime for their lightbulbs is 82 hours?

If the Norris lightbulbs showing 70 or fewer hours until burnout were discarded simply to provide an average lifetime of 82 hours, there is no question that discarding the lightbulbs with 70 or fewer hours until burnout is unethical. But, even if the discarded lightbulbs contain imperfections due to startup problems in the manufacturing process—and, as a result, should not have been included in the analysis—the statistician who conducted the study must account for all the data that were considered and explain how the sample actually used was obtained. To do otherwise is potentially misleading and would constitute unethical behavior on the part of both the company and the statistician.

A guideline in the shared values section of the American Statistical Association report states that statistical practitioners should avoid any tendency to slant statistical work toward predetermined outcomes. This type of unethical practice is often observed when unrepresentative samples are used to make claims. For instance, in many areas of the country smoking is not permitted in restaurants. Suppose, however, a lobbyist for the tobacco industry interviews people in restaurants where smoking is permitted in order to estimate the percentage of people who are in favor of allowing smoking in restaurants. The sample results show that 90% of the people interviewed are in favor of allowing smoking in restaurants. Based upon these sample results, the lobbyist claims that 90% of all people who eat in restaurants are in favor of permitting smoking in restaurants. In this case we would argue that sampling only persons eating in restaurants that allow smoking has biased the results. If only the final results of such a study are reported, readers unfamiliar with the details of the study (i.e., that the sample was collected only in restaurants allowing smoking) can be misled.

The scope of the American Statistical Association's report is broad and includes ethical guidelines that are appropriate not only for a statistician, but also for consumers of statistical information. We encourage you to read the report to obtain a better perspective of ethical issues as you continue your study of statistics and to gain the background for determining how to ensure that ethical standards are met when you start to use statistics in practice.

Summary

Statistics is the art and science of collecting, analyzing, presenting, and interpreting data. Nearly every college student majoring in business or economics is required to take a course in statistics. We began the chapter by describing typical statistical applications for business and economics.

Data consist of the facts and figures that are collected and analyzed. The four scales of measurement used to obtain data on a particular variable are nominal, ordinal, interval, and ratio. The scale of measurement for a variable is nominal when the data are labels or names used to identify an attribute of an element. The scale is ordinal if the data demonstrate the properties of nominal data and the order or rank of the data is meaningful. The scale is interval if the data demonstrate the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Finally, the scale of measurement is ratio if the data show all the properties of interval data and the ratio of two values is meaningful.

For purposes of statistical analysis, data can be classified as categorical or quantitative. Categorical data use labels or names to identify an attribute of each element. Categorical data use either the nominal or ordinal scale of measurement and may be nonnumeric or numeric. Quantitative data are numeric values that indicate how much or how many. Quantitative data use either the interval or ratio scale of measurement. Ordinary arithmetic operations are meaningful only if the data are quantitative. Therefore, statistical computations used for quantitative data are not always appropriate for categorical data.

In Sections 1.4 and 1.5 we introduced the topics of descriptive statistics and statistical inference. Descriptive statistics are the tabular, graphical, and numerical methods used to summarize data. The process of statistical inference uses data obtained from a sample to make estimates or test hypotheses about the characteristics of a population. The last four sections of the chapter provide information on the role of computers in statistical analysis, an introduction to the relatively new fields of analytics, data mining, and big data, and a summary of ethical guidelines for statistical practice.

Glossary

Analytics The scientific process of transforming data into insight for making better decisions.

Big data A set of data that cannot be managed, processed, or analyzed with commonly available software in a reasonable amount of time. Big data are characterized by great volume (a large amount of data), high velocity (fast collection and processing), or wide variety (could include nontraditional data such as video, audio, and text).

Categorical data Labels or names used to identify an attribute of each element. Categorical data use either the nominal or ordinal scale of measurement and may be nonnumeric or numeric.

Categorical variable A variable with categorical data.

Census A survey to collect data on the entire population.

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

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

Data mining The process of using procedures from statistics and computer science to extract useful information from extremely large databases.

Data set All the data collected in a particular study.

Descriptive analytics Analytical techniques that describe what has happened in the past.

Descriptive statistics Tabular, graphical, and numerical summaries of data.

Elements The entities on which data are collected.

Interval scale The scale of measurement for a variable if the data demonstrate the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numeric.

Nominal scale The scale of measurement for a variable when the data are labels or names used to identify an attribute of an element. Nominal data may be nonnumeric or numeric.

Observation The set of measurements obtained for a particular element.

Ordinal scale The scale of measurement for a variable if the data exhibit the properties of nominal data and the order or rank of the data is meaningful. Ordinal data may be nonnumeric or numeric.

Population The set of all elements of interest in a particular study.

Predictive analytics Analytical techniques that use models constructed from past data to predict the future or assess the impact of one variable on another.

Prescriptive analytics Analytical techniques that yield a course of action.

Quantitative data Numeric values that indicate how much or how many of something. Quantitative data are obtained using either the interval or ratio scale of measurement.

Quantitative variable A variable with quantitative data.

Ratio scale The scale of measurement for a variable if the data demonstrate all the properties of interval data and the ratio of two values is meaningful. Ratio data are always numeric.
Sample A subset of the population.
Sample survey A survey to collect data on a sample.
Statistical inference The process of using data obtained from a sample to make estimates or test hypotheses about the characteristics of a population.
Statistics The art and science of collecting, analyzing, presenting, and interpreting data.
Time series data Data collected over several time periods.
Variable A characteristic of interest for the elements.

Supplementary Exercises

1. Discuss the differences between statistics as numerical facts and statistics as a discipline or field of study.
2. Tablet PC Comparison provides a wide variety of information about tablet computers. Their website enables consumers to easily compare different tablets using factors such as cost, type of operating system, display size, battery life, and CPU manufacturer. A sample of 10 tablet computers is shown in Table 1.6 (Tablet PC Comparison website, February 28, 2013).
   a. How many elements are in this data set?
   b. How many variables are in this data set?
   c. Which variables are categorical and which variables are quantitative?
   d. What type of measurement scale is used for each of the variables?
3. Refer to Table 1.6.
   a. What is the average cost for the tablets?
   b. Compare the average cost of tablets with a Windows operating system to the average cost of tablets with an Android operating system.
   c. What percentage of tablets use a CPU manufactured by TI OMAP?
   d. What percentage of tablets use an Android operating system?
4. Table 1.7 shows data for eight cordless telephones (*Consumer Reports*, November 2012). The Overall Score, a measure of the overall quality for the cordless telephone, ranges from 0 to 100. Voice Quality has possible ratings of poor, fair, good, very good, and excellent. Talk Time is the manufacturer's claim of how long the handset can be used when it is fully charged.

TABLE 1.6 PRODUCT INFORMATION FOR 10 TABLET COMPUTERS

Tablet	Cost ($)	Operating System	Display Size (inches)	Battery Life (hours)	CPU Manufacturer
Acer Iconia W510	599	Windows	10.1	8.5	Intel
Amazon Kindle Fire HD	299	Android	8.9	9	TI OMAP
Apple iPad 4	499	iOS	9.7	11	Apple
HP Envy X2	860	Windows	11.6	8	Intel
Lenovo ThinkPad Tablet	668	Windows	10.1	10.5	Intel
Microsoft Surface Pro	899	Windows	10.6	4	Intel
Motorola Droid XYboard	530	Android	10.1	9	TI OMAP
Samsung Ativ Smart PC	590	Windows	11.6	7	Intel
Samsung Galaxy Tab	525	Android	10.1	10	Nvidia
Sony Tablet S	360	Android	9.4	8	Nvidia

TABLE 1.7 DATA FOR EIGHT CORDLESS TELEPHONES

Brand	Model	Price ($)	Overall Score	Voice Quality	Handset on Base	Talk Time (Hours)
AT&T	CL84100	60	73	Excellent	Yes	7
AT&T	TL92271	80	70	Very Good	No	7
Panasonic	4773B	100	78	Very Good	Yes	13
Panasonic	6592T	70	72	Very Good	No	13
Uniden	D2997	45	70	Very Good	No	10
Uniden	D1788	80	73	Very Good	Yes	7
Vtech	DS6521	60	72	Excellent	No	7
Vtech	CS6649	50	72	Very Good	Yes	7

a. How many elements are in this data set?

b. For the variables Price, Overall Score, Voice Quality, Handset on Base, and Talk Time, which variables are categorical and which variables are quantitative?

c. What scale of measurement is used for each variable?

5. Refer to the data set in Table 1.7.

a. What is the average price for the cordless telephones?

b. What is the average talk time for the cordless telephones?

c. What percentage of the cordless telephones have a voice quality of excellent?

d. What percentage of the cordless telephones have a handset on the base?

6. J.D. Power and Associates surveys new automobile owners to learn about the quality of recently purchased vehicles. The following questions were asked in the J.D. Power Initial Quality Survey, May 2012.

a. Did you purchase or lease the vehicle?

b. What price did you pay?

c. What is the overall attractiveness of your vehicle's exterior? (Unacceptable, Average, Outstanding, or Truly Exceptional)

d. What is your average number of miles per gallon?

e. What is your overall rating of your new vehicle? (l- to 10-point scale with 1 Unacceptable and 10 Truly Exceptional)

Comment on whether each question provides categorical or quantitative data.

7. The Kroger Company is one of the largest grocery retailers in the United States, with over 2000 grocery stores across the country. Kroger uses an online customer opinion questionnaire to obtain performance data about its products and services and learn about what motivates its customers (Kroger website, April 2012). In the survey, Kroger customers were asked if they would be willing to pay more for products that had each of the following four characteristics. The four questions were as follows:

Would you pay more for

products that have a brand name?
products that are environmentally friendly?
products that are organic?
products that have been recommended by others?

For each question, the customers had the option of responding Yes if they would pay more or No if they would not pay more.

a. Are the data collected by Kroger in this example categorical or quantitative?

b. What measurement scale is used?

8. *The Tennessean*, an online newspaper located in Nashville, Tennessee, conducts a daily poll to obtain reader opinions on a variety of current issues. In a recent poll, 762 readers