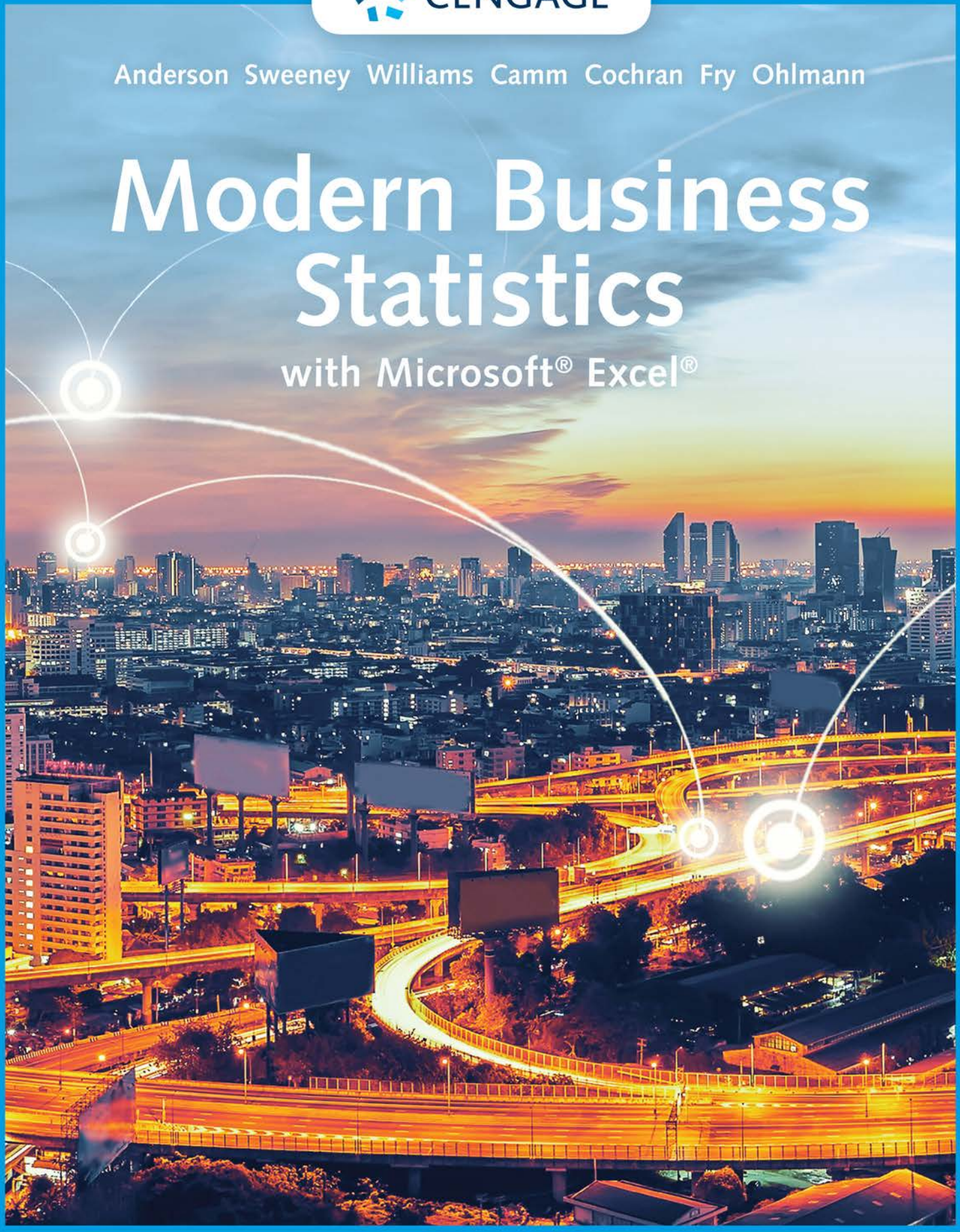
CENGAGE

Anderson Sweeney Williams Camm Cochran Fry Ohlmann

Modern Business Statistics

with Microsoft® Excel®



iStockPhoto.com/alongkot-s

Modern Business Statistics 7e

with Microsoft® 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
The University of Alabama

Michael J. Fry
University of Cincinnati

Jeffrey W. Ohlmann
University of Iowa

CENGAGE

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

Modern Business Statistics with Microsoft® Excel®, 7e
David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, Michael J. Fry, Jeffrey W. Ohlmann

Senior Vice President, Higher Education & Skills Product: Erin Joyner

Product Director: Jason Fremder

Senior Product Manager: Aaron Arnsparger

Content Manager: Conor Allen

Product Assistant: Maggie Russo

Marketing Manager: Chris Walz

Senior Learning Designer: Brandon Foltz

Associate Subject Matter Expert: Nancy Marchant

Digital Delivery Lead: Mark Hopkinson

Intellectual Property Analyst: Ashley Maynard

Intellectual Property Project Manager: Kelli Besse

Production Service: MPS Limited

Senior Project Manager, MPS Limited: Manoj Kumar

Art Director: Chris Doughman

Text Designer: Beckmeyer Design

Cover Designer: Beckmeyer Design

Cover Image: iStockPhoto.com/alongkot-s

Library of Congress Control Number: 2019915730

Student Edition:
ISBN: 978-0-357-13138-1

Loose-leaf Edition:
ISBN: 978-0-357-13139-8

Cengage
200 Pier 4 Boulevard
Boston, MA 02210
USA

Printed in the United States of America
Print Number: 01 Print Year: 2019

Brief Contents

Contents

Preface

This text is the seventh edition of *Modern Business Statistics with Microsoft® Excel®*. With this edition we welcome two eminent scholars to our author team: Michael J. Fry of the University of Cincinnati and Jeffrey W. Ohlmann of the University of Iowa. Both Mike and Jeff are accomplished teachers, researchers, and practitioners in the fields of statistics and business analytics. You can read more about their accomplishments in the About the Authors section that follows this preface. We believe that the addition of Mike and Jeff as our coauthors will both maintain and improve the effectiveness of *Modern Business Statistics with Microsoft Excel*.

The purpose of *Modern Business Statistics with Microsoft 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 to decisions 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

Modern Business Statistics with Microsoft 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 *Modern Business Statistics with Microsoft 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, and copying we do not assume that readers are familiar with Excel or Excel's tools for statistical analysis. As a result, we have included Appendix E, which provides an introduction to Excel 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 use 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. 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 *Modern Business Statistics with Microsoft 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.

- **Software.** In addition to step-by-step instructions and screen captures that show how to use the latest version of Excel to implement statistical procedures, we also provide instructions for Excel Online and R through the MindTap Reader.
- **New Examples and Exercises Based on Real Data.** In this edition, we have added headers to all Applications exercises to make the application of each exercise more clear. We have also added over 160 new examples and exercises based on real data and referenced sources. By using data from sources also used by *The Wall Street Journal*, *USA Today*, *The Financial Times*, *Forbes*, and others, we have drawn from actual studies and applications to develop explanations and create exercises that demonstrate the many uses of statistics in business and economics. We believe that the use of real data from interesting and relevant problems generates greater student interest in the material and enables the student to more effectively learn about both statistical methodology and its application.
- **Case Problems.** We have added four new case problems to this edition. The 47 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.
- **Appendixes for Use of R.** We now provide appendixes in the MindTap Reader for many chapters that demonstrate the use of the popular open-source software R and RStudio for statistical applications. The use of R is not required to solve any problems or cases in the textbook, but the appendixes provide an introduction to R and RStudio for interested instructors and students.

Features and Pedagogy

Authors Anderson, Sweeney, Williams, Camm, Cochran, Fry, and Ohlmann have continued many of the features that appeared in previous editions. Important ones for students are noted here.

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.

Margin Annotations and Notes and Comments

Margin annotations that highlight key points and provide additional insights for the student are a key feature of this text. These annotations, which appear in the margins, 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

Over 250 data files are available on the website that accompanies the text. 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. The R appendixes for many of the chapters in the text are also accessible through MindTap.

MindTap is a customizable digital course solution that includes an interactive eBook and autograded, algorithmic exercises from the textbook. 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 at www.cengage.com/decisionsciences/anderson/mbs/7e.

For Instructors

Instructor resources are available to adopters on the Instructor Companion Site, which can be found and accessed at www.cengage.com/decisionsciences/anderson/mbs/7e, 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 in 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.

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 senior product manager, Aaron Arnsparger; our content manager, Conor Allen; senior learning designer, Brandon Foltz; digital delivery lead, Mark Hopkinson; and our senior project managers at MPS Limited, Santosh Pandey & Manoj Kumar, 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:

Jamal Abdul-Hafidh
University of Missouri–St. Louis

Chris Adalikwu
Concordia College

Eugene Allevato
Woodbury University

Solomon Antony
Murray State University

Ardavan Asef-Vaziri
California State University, Northridge

S. Scott Bailey
Troy University

Robert J. Banis
University of Missouri–St. Louis

Wayne Bedford
University of West Alabama

Enoch K. Beraho
South Carolina State University

Timothy M. Bergquist
Northwest Christian College

Darl Bien
University of Denver

William H. Bleuel
Pepperdine University

Gary Bliss
Florida State University–Panama City

Leslie M. Bobb
New York Institute of Technology

Michelle Boddy
Baker College

Thomas W. Bolland
Ohio University

Derrick S. Boone, Sr.
Wake Forest University

Lawrence Bos
Cornerstone University

Alan Brokaw
Michigan Tech University

Nancy Brooks
University of Vermont

Yvonne Brown
Pima Community College

Dawn Bulriss
Maricopa Community Colleges

Robert Burgess
Georgia Tech

Von L. Burton
Athens State University

John R. Carpenter
Cornerstone University

Jasmine Chang
Georgia State University

Si Chen
Murray State University

Alan S. Chesen
Wright State University

Michael Cicero
Highline Community College

Robert Collins
Marquette University

Ping Deng
Maryville University

Sarvanan Devaraj
Notre Dame University

Terry Dielman
Texas Christian University

Cassandra DiRienzo
Elon University

Anne Drougas
Dominican University

Jianjun Du
University of Houston, Victoria

John N. Dyer
Georgia Southern University

Hossein Eftekari
University of Wisconsin–River Falls

Mohammed A. El-Saidi
Ferris State University

Robert M. Escudero
Pepperdine University

Allessandra Faggian
Ohio State University

Nicolas Farnum
California State University, Fullerton

Abe Feinberg
California State University, Northridge

Maggie Williams Flint
Northeast State Tech Community College

Alfonso Flores-Lagunes
University of Arizona

James Flynn
Cleveland State University

Alan F. Foltz
Drury University

Ronald L. Friesen
Bluffton College

Richard Gebhart
University of Tulsa

Paul Gentine
Bethany College

Deborah J. Gougeon
University of Scranton

Jeffrey Gropp
DePauw University

V. Daniel Guide
Duquesne University

Aravind Narasipur
Chennai Business School

Rhonda Hensley
North Carolina A&T University

Erick Hofacker
University of Wisconsin–River Falls

Amy C. Hooper
Gettysburg College

Paul Hudec
Milwaukee School of Engineering

Alan Humphrey
University of Rhode Island

Wade Jackson
University of Memphis

Timmy James
Northwest Shoals Community College

Eugene Jones
The Ohio State University

Naser Kamleh
Wallace Community College

Mark P. Karscig
Central Missouri State University

Howard Kittleson
Riverland Community College

Kenneth Klassen
California State University, Northridge

Eileen Quinn Knight
St. Xavier University, Chicago

Bharat Kolluri
University of Hartford

Joseph Kosler
Indiana University of Pennsylvania

David A. Kravitz
George Mason University

Laura Kuhl
University of Phoenix, Cleveland Campus

June Lapidus
Roosevelt University

John Lawrence
California State University, Fullerton

Tenpao Lee
Niagara University

Daniel Light
Northwest State College

Robert Lindsey
College of Charleston

B. Lucas
A&M College

Michael Machiorlatti
City College of San Francisco

Malik B. Malik
University of Maryland Eastern Shore

Lee McClain
Western Washington University

Timothy E. McDaniel
Buena Vista University

Kim I. Melton
North Georgia College & State University

Brian Metz
Cabrini College

John M. Miller
Sam Houston State University

Patricia A. Mullins
University of Wisconsin–Madison

Jack Muryn
University of Wisconsin, Washington County

Muhammad Mustafa
South Carolina State University

Anthony Narsing
Macon State College

Kenneth F. O'Brien
Farmingdale State College

Ceyhun Ozgur
Valparaiso University

Michael Parzen
Emory University

Barry Pasternack
California State University, Fullerton

Lynne Pastor
Carnegie Mellon University

Ranjna Patel
Bethune-Cookman College

Tremaine Pimperl
Faulkner State Community College

Jennifer M. Platania
Elon University

Von Roderick Plessner
Northwest State Community College

Glenn Potts
University of Wisconsin–River Falls

Irene Powell
Grinnell College

Leonard Presby
William Paterson University

W. N. Pruitt
South Carolina State University

Narseeyappa Rajanikanth
Mississippi Valley State University

Elizabeth L. Rankin
Centenary College of Louisiana

Tim Raynor
Albertus Magnus College

Carolyn Renier
Pellissippi State Tech Community College

Ronny Richardson
Southern Polytechnic State University

Leonard E. Ross
California State University, Pomona

Probir Roy
University of Missouri, Kansas City

Randall K. Russell
Yavapai College

Alan Safer
California State University, Long Beach

David Satava
University of Houston, Victoria

Richard W. Schrader
Bellarmine University

Larry Seifert
Webster University

John Seydel
Arkansas State University

Jim Shi
Robinson College of Business
Georgia State University

Philip Shaw
Fairfield University

Robert Simoneau
Keene State College

Harvey A. Singer
George Mason University

Donald R. Smith
Monmouth University

Toni M. Somers
Wayne State University

Clifford Sowell
Berea College

Keith Starcher
Indiana Wesleyan University

William Stein
Texas A&M University

Jason Stine
Troy University

William Struning
Seton Hall University

Timothy S. Sullivan
Southern Illinois University, Edwardsville

Peter Wibawa Sutanto
Prairie View A&M University

Lee Tangedahk
University of Montana

Sudhir Thakur
California State University, Sacramento

Alexander Thomson
Schoolcraft College

Suzanne Tilleman
Montana State University Northern

Daniel Tschopp
Daemen College (NY)

Sushila Umashankar
University of Arizona

Jack Vaughn
University of Texas, El Paso

Dave Vinson
Pellissippi State Community College

John Vogt
Newman University

Geoffrey L. Wallace
University of Wisconsin, Madison

Michael Wiemann
Metro Community Colleges

Charles Wilf
Duquesne University

John Wiorkowski
University of Texas, Dallas

Joyce A. Zadzilka
Morrisville State College

Guoqiang Peter Zhang
Georgia Southern University

Zhe George Zhang
Western Washington University

Deborah G. Ziegler
Hannibal-LaGrange College

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

Mike Bourke
Houston Baptist University

Peter Bryant
University of Colorado

Terri L. Byczkowski
University of Cincinnati

Ying Chien
University of Scranton

Robert Cochran
University of Wyoming

Murray Côté
University of Florida

David W. Cravens
Texas Christian University

Robert Carver
Stonehill College

Tom Dahlstrom
Eastern College

Ronald Ehresman
Baldwin-Wallace College

Michael Ford
Rochester Institute of Technology

Phil Fry
Boise State University

Paul Guy
California State University, Chico

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

About the Authors

David R. Anderson. David R. Anderson is Professor Emeritus in the Carl H. Lindner College of Business at the University of Cincinnati. Born in Grand Forks, North Dakota, he earned his B.S., M.S., and Ph.D. degrees from Purdue University. Professor Anderson has served as Head of the Department of Quantitative Analysis and Operations Management and as Associate Dean of the College of Business Administration at the University of Cincinnati. In addition, he was the coordinator of the College's first Executive Program.

At the University of Cincinnati, Professor Anderson has taught introductory statistics for business students as well as graduate-level courses in regression analysis, multivariate analysis, and management science. He has also taught statistical courses at the Department of Labor in Washington, D.C. He has been honored with nominations and awards for excellence in teaching and excellence in service to student organizations.

Professor Anderson has coauthored 10 textbooks in the areas of statistics, management science, linear programming, and production and operations management. He is an active consultant in the field of sampling and statistical methods.

Dennis J. Sweeney. Dennis J. Sweeney is Professor Emeritus in the Carl H. Lindner College of Business at the University of Cincinnati. Born in Des Moines, Iowa, he earned a B.S.B.A. degree from Drake University and his M.B.A. and D.B.A. degrees from Indiana University, where he was an NDEA Fellow. Professor Sweeney has worked in the management science group at Procter & Gamble and spent a year as a visiting professor at Duke University. Professor Sweeney served as Head of the Department of Quantitative Analysis and as Associate Dean of the College of Business Administration at the University of Cincinnati.

Professor Sweeney has published more than 30 articles and monographs in the area of management science and statistics. The National Science Foundation, IBM, Procter & Gamble, Federated Department Stores, Kroger, and Duke Energy have funded his research, which has been published in *Management Science*, *Operations Research*, *Mathematical Programming*, *Decision Sciences*, and other journals.

Professor Sweeney has coauthored 10 textbooks in the areas of statistics, management science, linear programming, and production and operations management.

Thomas A. Williams. Thomas A. Williams is Professor Emeritus of Management Science in the College of Business at Rochester Institute of Technology. Born in Elmira, New York, he earned his B.S. degree at Clarkson University. He did his graduate work at Rensselaer Polytechnic Institute, where he received his M.S. and Ph.D. degrees.

Before joining the College of Business at RIT, Professor Williams served for seven years as a faculty member in the College of Business Administration at the University of Cincinnati, where he developed the undergraduate program in Information Systems and then served as its coordinator. At RIT he was the first chairman of the Decision Sciences Department. He has taught courses in management science and statistics, as well as graduate courses in regression and decision analysis.

Professor Williams is the coauthor of 11 textbooks in the areas of management science, statistics, production and operations management, and mathematics. He has been a consultant for numerous Fortune 500 companies and has worked on projects ranging from the use of data analysis to the development of large-scale regression models.

Jeffrey D. Camm. Jeffrey D. Camm is the Inmar Presidential Chair and Associate Dean of Business Analytics in the School of Business at Wake Forest University. Born in Cincinnati, Ohio, he holds a B.S. from Xavier University (Ohio) and a Ph.D. from Clemson University.

Prior to joining the faculty at Wake Forest, he was on the faculty of the University of Cincinnati. He has also been a visiting scholar at Stanford University and a visiting professor of business administration at the Tuck School of Business at Dartmouth College.

Dr. Camm has published over 30 papers in the general area of optimization applied to problems in operations management and marketing. He has published his research in *Science*, *Management Science*, *Operations Research*, *Interfaces*, and other professional journals. Dr. Camm was named the Dornoff Fellow of Teaching Excellence at the University of Cincinnati and he was the 2006 recipient of the INFORMS Prize for the Teaching of Operations Research Practice. A firm believer in practicing what he preaches, he has served as an operations research consultant to numerous companies and government agencies. From 2005 to 2010 he served as editor-in-chief of the *INFORMS Journal on Applied Analytics* (formerly *Interfaces*).

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

Professor Cochran has published over 40 papers in the development and application of operations research and statistical methods. He has published his research in *Management Science*, *The American Statistician*, *Communications in Statistics—Theory and Methods*, *Annals of Operations Research*, *European Journal of Operational Research*, *Journal of Combinatorial Optimization*, *Interfaces*, *Statistics and Probability Letters*, and other professional journals. He was the 2008 recipient of the INFORMS Prize for the Teaching of Operations Research Practice and the 2010 recipient of the Mu Sigma Rho Statistical Education Award. Professor Cochran was elected to the International Statistics Institute in 2005, was named a Fellow of the American Statistical Association in 2011, and was named a Fellow of INFORMS in 2017. He also received the Founders Award in 2014 and the Karl E. Peace Award in 2015 from the American Statistical Association as well as the President's Award in 2018 from INFORMS. A strong advocate for effective operations research and statistics education as a means of improving the quality of applications to real problems, Professor Cochran has organized and chaired teaching effectiveness workshops in Uruguay, South Africa, Colombia, India, Argentina, Kenya, Cuba, Croatia, Cameroon, Nepal, Moldova, and Bulgaria. He has served as a statistical consultant to numerous companies and not-for-profit organizations. He served as editor-in-chief of *INFORMS Transactions on Education* from 2006 to 2012 and is on the editorial board of *INFORMS Journal of Applied Analytics* (formerly *Interfaces*), *International Transactions in Operational Research*, and *Significance*.

Michael J. Fry. Michael J. Fry is Professor of Operations, Business Analytics, and Information Systems (OBAIS) and Academic Director of the Center for Business Analytics in the Carl H. Lindner College of Business at the University of Cincinnati. Born in Killeen, Texas, he earned a B.S. from Texas A&M University, and M.S.E. and Ph.D. degrees from the University of Michigan. He has been at the University of Cincinnati since 2002, where he was previously Department Head and has been named a Lindner Research Fellow. He has also been a visiting professor at the Samuel Curtis Johnson Graduate School of Management at Cornell University and the Sauder School of Business at the University of British Columbia.

Professor Fry has published more than 25 research papers in journals such as *Operations Research*, *M&SOM*, *Transportation Science*, *Naval Research Logistics*, *IIE Transactions*, *Critical Care Medicine*, and *Interfaces*. He serves on editorial boards for journals such as *Production and Operations Management*, *INFORMS Journal of Applied Analytics* (formerly *Interfaces*), *Omega*, and *Journal of Quantitative Analysis in Sports*. His research interests are in applying analytics to the areas of supply chain management, sports, and public-policy

operations. He has worked with many different organizations for his research, including Dell, Inc., Starbucks Coffee Company, Great American Insurance Group, the Cincinnati Fire Department, the State of Ohio Election Commission, the Cincinnati Bengals, and the Cincinnati Zoo & Botanical Garden. He was named a finalist for the Daniel H. Wagner Prize for Excellence in Operations Research Practice, and he has been recognized for both his research and teaching excellence at the University of Cincinnati. In 2019, he led the team that was awarded the INFORMS UPS George D. Smith Prize on behalf of the OBAIS Department at the University of Cincinnati.

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

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

Chapter 1

Data and Statistics

CONTENTS

STATISTICS IN PRACTICE

Bloomberg Businessweek*

NEW YORK, NEW YORK

Bloomberg Businessweek is one of the most widely read business magazines in the world. 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* 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 North American subscriber survey indicated 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 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.

Bloomberg Businessweek *uses statistical facts and summaries in many of its articles. AP Images/Weng lei-Imaginechina*

*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:

- Unemployment dropped to an 18-year low of 3.8% in May 2018 from 3.9% in April and after holding at 4.1% for the prior six months (*Wall Street Journal,* June 1, 2018).
- Tesla ended 2017 with around $5.4 billion of liquidity. Analysts forecast it will burn through $2.8 billion of cash this year (*Bloomberg Businessweek,* April 19, 2018).
- The biggest banks in America reported a good set of earnings for the first three months of 2018. Bank of America and Morgan Stanley made quarterly net profits of $6.9 billion and $2.7 billion, respectively (*The Economist,* April 21, 2018).
- According to a study from the Pew Research Center, 15% of U.S. adults say they have used online dating sites or mobile apps (*Wall Street Journal,* May 2, 2018).

- According to the U.S. Centers for Disease Control and Prevention, in the United States alone, at least 2 million illnesses and 23,000 deaths can be attributed each year to antibiotic-resistant bacteria (*Wall Street Journal,* February 13, 2018).

The numerical facts in the preceding statements—3.8%, 3.9%, 4.1%, $5.4 billion, $2.8 billion, $6.9 billion, $2.7 billion, 15%, 2 million, 23,000—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 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 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 average dividend yield for the S&P 500 companies for 2017 was 1.88%. Over the same period the average dividend yield for Microsoft was 1.72% (*Yahoo Finance*). In this case, the statistical information on dividend yield indicates a lower 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 The Nielsen Company and IRI 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 as an observer.
- Per Capita Gross Domestic Product (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.
- 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, 3615, BB–, and Stable. The second observation contains the following measurements: Member, 49755, 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

[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.

DATA*file*
Nations

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

TABLE 1.1 Data Set for 60 Nations in the World Trade Organization

Nation	WTO Status	Per Capita GDP ($)	Fitch Rating	Fitch Outlook
Armenia	Member	3,615	BB−	Stable
Australia	Member	49,755	AAA	Stable
Austria	Member	44,758	AAA	Stable
Azerbaijan	Observer	3,879	BBB−	Stable
Bahrain	Member	22,579	BBB	Stable
Belgium	Member	41,271	AA	Stable
Brazil	Member	8,650	BBB	Stable
Bulgaria	Member	7,469	BBB−	Stable
Canada	Member	42,349	AAA	Stable
Cape Verde	Member	2,998	B+	Stable
Chile	Member	13,793	A+	Stable
China	Member	8,123	A+	Stable
Colombia	Member	5,806	BBB−	Stable
Costa Rica	Member	11,825	BB+	Stable
Croatia	Member	12,149	BBB−	Negative
Cyprus	Member	23,541	B	Negative
Czech Republic	Member	18,484	A+	Stable
Denmark	Member	53,579	AAA	Stable
Ecuador	Member	6,019	B−	Positive
Egypt	Member	3,478	B	Negative
El Salvador	Member	4,224	BB	Negative
Estonia	Member	17,737	A+	Stable
France	Member	36,857	AAA	Negative
Georgia	Member	3,866	BB−	Stable
Germany	Member	42,161	AAA	Stable
Hungary	Member	12,820	BB+	Stable
Iceland	Member	60,530	BBB	Stable
Ireland	Member	64,175	BBB+	Stable
Israel	Member	37,181	A	Stable
Italy	Member	30,669	A−	Negative
Japan	Member	38,972	A+	Negative
Kazakhstan	Observer	7,715	BBB+	Stable
Kenya	Member	1,455	B+	Stable
Latvia	Member	14,071	BBB	Positive
Lebanon	Observer	8,257	B	Stable
Lithuania	Member	14,913	BBB	Stable
Malaysia	Member	9,508	A−	Stable
Mexico	Member	8,209	BBB	Stable
Peru	Member	6,049	BBB	Stable
Philippines	Member	2,951	BB+	Stable
Poland	Member	12,414	A−	Positive
Portugal	Member	19,872	BB+	Negative
South Korea	Member	27,539	AA−	Stable
Romania	Member	9,523	BBB−	Stable
Russia	Member	8,748	BBB	Stable
Rwanda	Member	703	B	Stable
Serbia	Observer	5,426	BB−	Negative
Singapore	Member	52,962	AAA	Stable
Slovakia	Member	16,530	A+	Stable

Slovenia	Member	21,650	A−	Negative
South Africa	Member	5,275	BBB	Stable
Spain	Member	26,617	A−	Stable
Sweden	Member	51,845	AAA	Stable
Switzerland	Member	79,888	AAA	Stable
Thailand	Member	5,911	BBB	Stable
Turkey	Member	10,863	BBB−	Stable
United Kingdom	Member	40,412	AAA	Negative
United States	Member	57,638	AAA	Stable
Uruguay	Member	15,221	BB+	Positive
Zambia	Member	1,270	B+	Negative

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

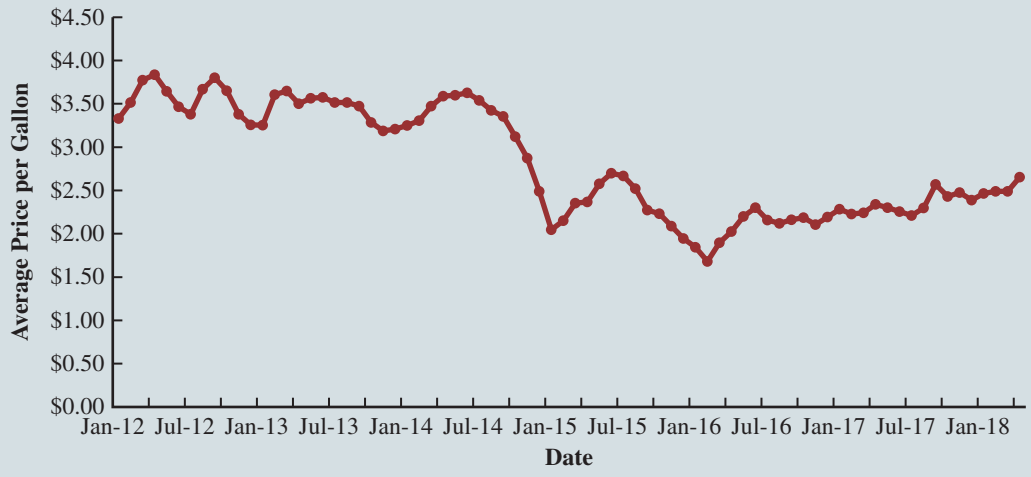
For purposes of statistical analysis, distinguishing between cross-sectional data and time series data is important. **Cross-sectional data** are data collected at the same or approximately the same point in time. The data in Table 1.1 are cross-sectional because they describe the five variables for 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 2012 and 2018. From January 2012 until June 2014, prices fluctuated between $3.19 and $3.84 per gallon before a long stretch of decreasing prices from July 2014 to January 2015. The lowest average price per gallon occurred in January 2016 ($1.68). Since then, the average price appears to be on a gradual increasing trend.

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 2008 to 2018. Poor economic conditions caused a serious drop in the index during 2008 with the low point occurring in February 2009 (7,062). After that, the index has been on a remarkable nine-year increase, reaching its peak (26,149) in January 2018.

The graph in Panel (B) shows the net income of McDonald's Inc. from 2008 to 2017. 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, decreased slightly in 2012, and peaked in 2013. After three years of relatively lower net income, their net income increased to $5.19 billion in 2017.

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

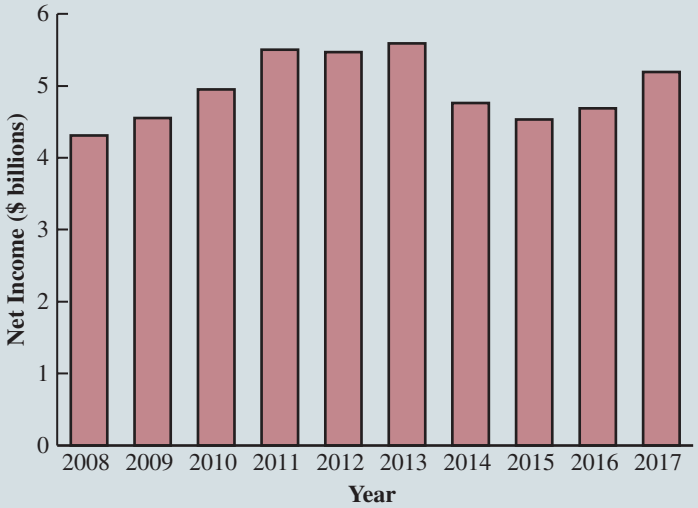
FIGURE 1.1 U.S. Average Price per Gallon for Conventional Regular Gasoline



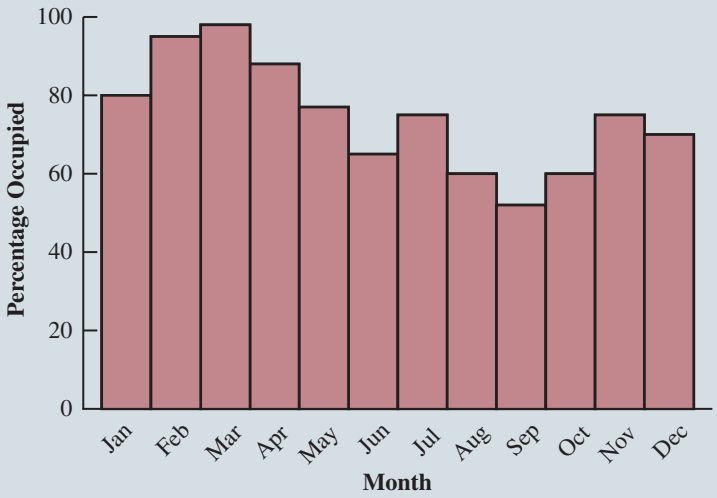
Source: *Energy Information Administration, U.S. Department of Energy.*

FIGURE 1.2 A Variety of Graphs of Time Series Data

(A) Dow Jones Industrial Average Index



(B) Net Income for McDonald's Inc.



(C) Occupancy Rate of South Florida Hotels

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 + 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.

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. The Nielsen Company and IRI built successful businesses collecting and processing data that they sell to advertisers and product manufacturers.

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

Data are also available from a variety of industry associations and special interest organizations. The U.S. Travel Association 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 website DATA.GOV was launched by the U.S. government in 2009 to make it easier for the public to access data collected by the U.S. federal government. The DATA.GOV website includes over 150,000 data sets from a variety of U.S. federal departments and agencies, but there are many other federal agencies that maintain their own websites and data repositories. Table 1.3 lists selected governmental agencies and some of the data they provide. Figure 1.3 shows the home-page for the DATA.GOV website. Many state and local governments are also now providing data sets online. As examples, the states of California and Texas maintain open data portals at data.ca.gov and data.texas.gov, respectively. New York City's open data website is opendata.cityofnewyork.us and the city of Cincinnati, Ohio, is at data.cincinnati-oh.gov.

Observational Study

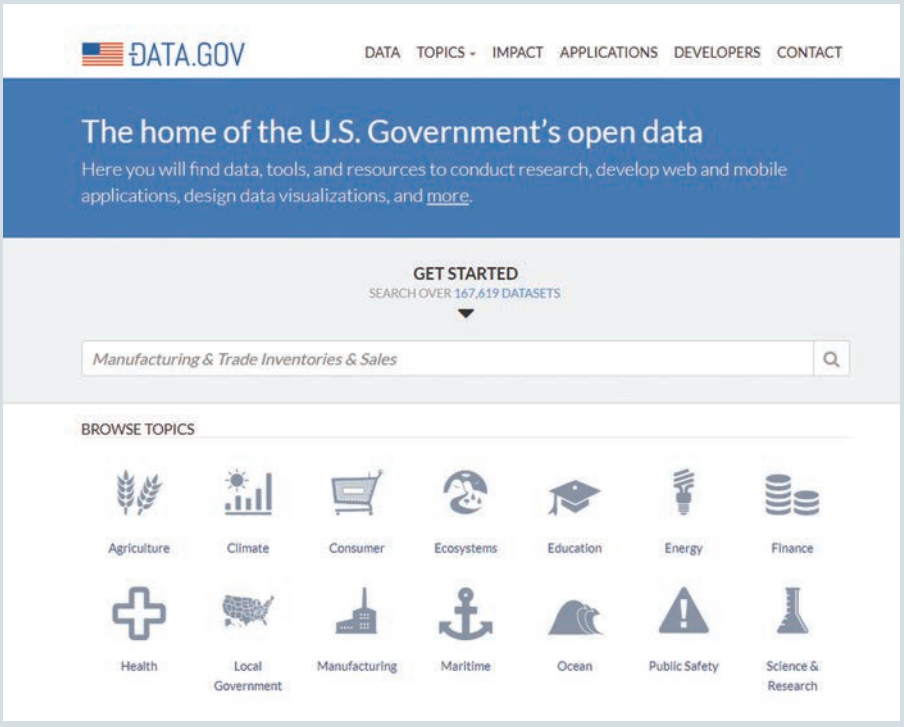
In an *observational study* we simply observe what is happening in a particular situation, record data on one or more variables of interest, and conduct a statistical analysis of the resulting data. For example, researchers might observe a randomly selected group of customers that enter a Walmart supercenter to collect data on variables such as the length of time the customer spends shopping, the gender of the customer, and the amount spent. 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

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
DATA.GOV	More than150,000 data sets including agriculture, consumer, education, health, and manufacturing data

FIGURE 1.3 DATA.GOV Homepage



U.S. BUREAU OF LABOR STATISTICS HOMEPAGE

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.

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.

Experiment

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.

In Chapter 13 we discuss statistical methods appropriate for analyzing the data from an experiment.

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.

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 the media, 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

TABLE 1.4 Frequencies and Percent Frequencies for the Fitch Credit Rating Outlook of 60 Nations

Fitch Outlook	Frequency	Percent Frequency (%)
Positive	4	6.7
Stable	44	73.2
Negative	12	20.0

FIGURE 1.4 Bar Chart for the Fitch Credit Rating Outlook for 60 Nations

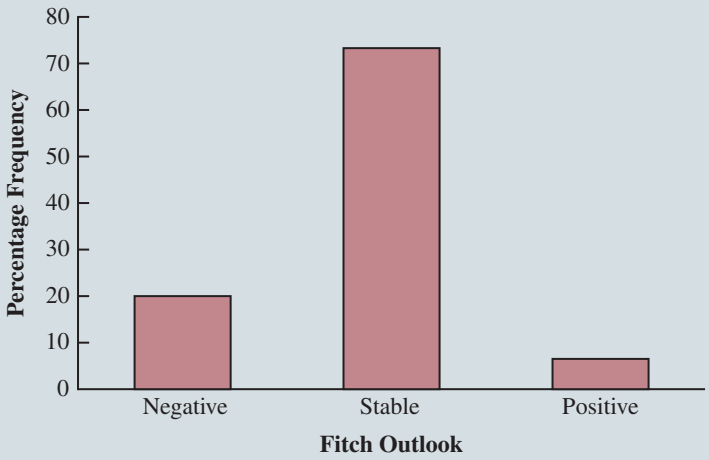
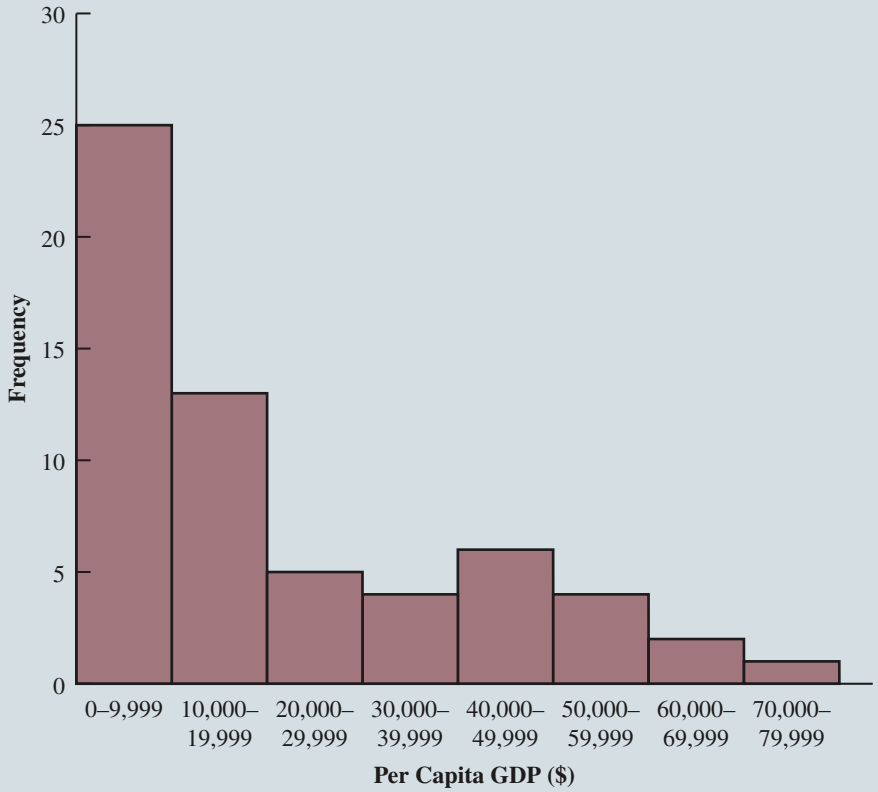


Figure 1.4, we can see that the majority of Fitch Outlook credit ratings are stable, with 73.2% of the nations having this rating. More nations have a negative outlook (20%) than a positive outlook (6.7%).

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

FIGURE 1.5 Histogram of Per Capita GDP for 60 Nations



to see that Per Capita GDP for the 60 nations ranges from $0 to $80,000, with the highest concentration between $0 and $10,000. Only one nation had a Per Capita GDP exceeding $70,000.

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 $21,279. This average provides a measure of the central tendency, or central location of the data.

Chapters 2 and 3 devote attention to the tabular, graphical, and numerical methods of descriptive statistics.

There is a great deal of interest in effective methods for developing and presenting 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 Rogers Industries. Rogers manufactures lithium batteries used in rechargeable electronics such as laptop computers and tablets. In an attempt to increase battery life for its products, Rogers has developed a new solid-state lithium battery that should last longer and be safer to use. In this case, the population is defined as all lithium batteries that could be produced using the new solid-state technology. To evaluate the advantages of the new battery, a sample of 200 batteries manufactured with the new solid-state technology were tested. Data collected from this sample showed the number of hours each battery lasted before needing to be recharged under controlled conditions. See Table 1.5.

Suppose Rogers wants to use the sample data to make an inference about the average hours of battery life for the population of all batteries that could be produced with the new solid-state technology. Adding the 200 values in Table 1.5 and dividing the total by 200 provides the sample average battery life: 18.84 hours. We can use this sample result to estimate that the average lifetime for the batteries in the population is 18.84 hours. Figure 1.6 provides a graphical summary of the statistical inference process for Rogers Industries.

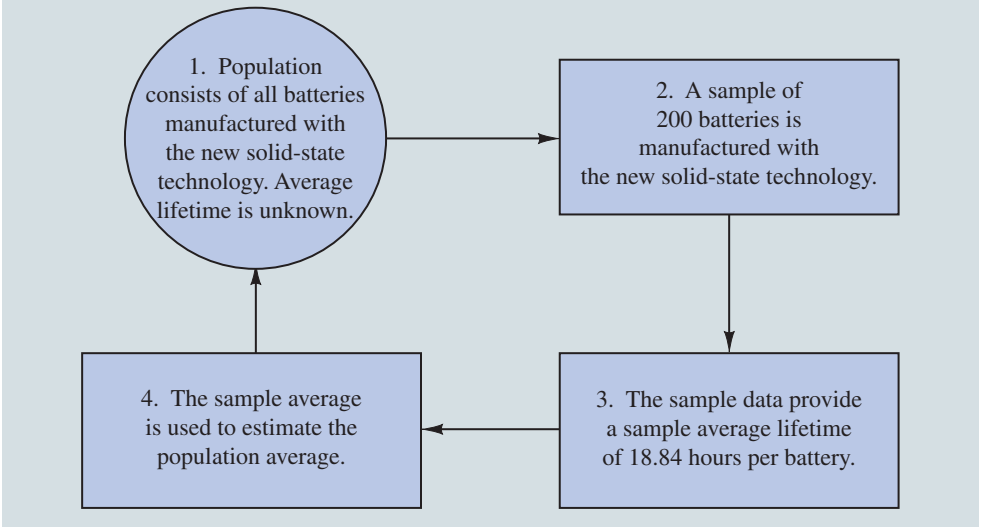
Whenever statisticians use a sample to estimate a population characteristic of interest, they usually provide a statement of the quality, or precision, associated with the estimate. For the Rogers Industries example, the statistician might state that the point estimate of the average battery life is 18.84 hours $\pm$.68 hours. Thus, an interval estimate of the average battery life is 18.16 to 19.52 hours. The statistician can also state how confident he or she is that the interval from 18.16 to 19.52 hours contains the population average.

DATA*file*
Rogers

TABLE 1.5 Hours Until Recharge for a Sample of 200 Batteries for the Rogers Industries Example

Battery Life (hours)									
19.49	18.18	18.65	19.45	19.89	18.94	17.72	18.35	18.66	18.23
19.08	19.92	19.01	18.84	17.73	19.70	18.37	18.69	19.98	18.80
19.11	18.26	19.05	17.89	19.61	18.52	18.10	19.08	18.27	18.29
19.55	18.81	18.68	17.43	20.34	17.73	17.66	18.52	19.90	19.33
18.81	19.12	18.39	19.27	19.43	19.29	19.11	18.96	19.65	18.20
19.18	20.07	18.54	18.37	18.13	18.29	19.11	20.22	18.07	18.91
18.44	19.04	18.88	19.51	18.84	20.98	18.82	19.40	19.00	17.53
18.74	19.04	18.35	19.01	17.54	18.14	19.82	19.23	19.20	20.02
20.14	17.75	18.50	19.85	18.93	19.07	18.83	18.54	17.85	18.51
18.74	18.74	19.06	19.00	18.77	19.12	19.58	18.75	18.67	20.71
18.35	19.42	19.42	19.41	19.85	18.23	18.31	18.44	17.61	19.21
17.71	18.04	19.53	18.87	19.11	19.28	18.55	18.58	17.33	18.75
18.52	19.06	18.54	18.41	19.86	17.24	18.32	19.27	18.34	18.89
18.78	18.88	18.67	18.19	19.07	20.12	17.69	17.92	19.49	19.52
19.91	18.46	18.98	19.18	19.01	18.79	17.90	18.43	18.35	19.02
18.06	19.11	19.40	18.71	18.91	18.95	18.51	19.27	20.39	19.72
17.48	17.49	19.29	18.49	17.93	19.42	19.19	19.46	18.56	18.41
18.24	17.83	18.28	19.51	18.17	18.64	18.57	18.65	18.61	17.97
18.73	19.32	19.37	18.60	19.16	19.44	18.28	19.20	17.88	18.90
19.66	19.00	18.43	19.54	19.15	18.62	19.64	18.87	18.31	19.54

FIGURE 1.6 The Process of Statistical Inference for the Rogers Industries Example



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

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.*

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:El contain the variable names; cells A2:A61 contain the observation names; and cells B2:E61 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.

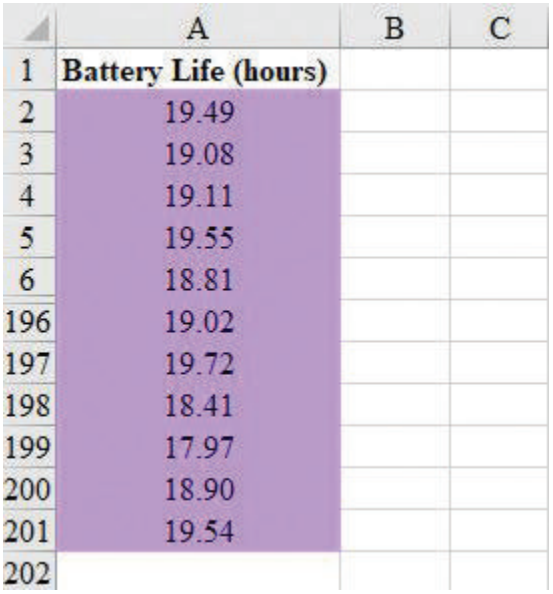
The data are the focus of the statistical analysis. Except for the headings in row 1, each row of the worksheet corresponds to an observation and each column corresponds to a variable. For instance, row 2 of the worksheet contains the data for the first observation, Armenia; row 3 contains the data for the second observation, Australia; row 3 contains the data for the third observation, Austria; and so on. The names in column A provide a

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
1	**Nation**	**WTO Status**	**Per Capita GDP ($)**	**Fitch Rating**	**Fitch Outlook**	
2	Armenia	Member	3615	BB-	Stable	
3	Australia	Member	49755	AAA	Stable	
4	Austria	Member	44758	AAA	Stable	
5	Azerbaijan	Observer	3879	BBB-	Stable	
6	Bahrain	Member	22579	BBB	Stable	
7	Belgium	Member	41271	AA	Stable	
8	Brazil	Member	8650	BBB	Stable	
9	Bulgaria	Member	7469	BBB-	Stable	
10	Canada	Member	42349	AAA	Stable	
11	Cape Verde	Member	2998	B+	Stable	
12	Chile	Member	13793	A+	Stable	
13	China	Member	8123	A+	Stable	
14	Colombia	Member	5806	BBB-	Stable	
55	Switzerland	Member	79888	AAA	Stable	
56	Thailand	Member	5911	BBB	Stable	
57	Turkey	Member	10863	BBB-	Stable	
58	United Kingdom	Member	40412	AAA	Negative	
59	Uruguay	Member	15221	BB+	Positive	
60	United Stabletes	Member	57638	AAA	Stable	
61	Zambia	Member	1270	B+	Negative	
62						

FIGURE 1.8 Excel Worksheet for the Rogers Industries Data Set



	A	B	C
1	Battery Life (hours)		
2	19.49		
3	19.08		
4	19.11		
5	19.55		
6	18.81		
196	19.02		
197	19.72		
198	18.41		
199	17.97		
200	18.90		
201	19.54		
202			

DATA*file*
Rogers

Note: Rows 7–195 are hidden

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 Rogers Industries 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. 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 Chart, and Using Excel to Construct a Frequency Distribution. 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 files 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, and data labels.

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 batteries 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 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.

DATA *file*
Rogers

Enter/Access Data: Open the file *Rogers*. The data are in cells A2:A201 and the label is in cell A1.

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

$$=AVERAGE(A2:A201)$$

Similarly, the formula $=MEDIAN(A2:A201)$ is entered in cell E3 to compute the median.

To identify the result, the label "Average Lifetime" is entered in cell D2 and the label "Median Lifetime" is entered into cell D3. 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 18.84 hours and that the value computed using the MEDIAN function is also 18.84 hours.

FIGURE 1.9 Computing the Average Lifetime of Batteries for Rogers Industries Using Excel's Average Function

	A	B	C	D	E	F
1	**Battery Life (hours)**					
2	19.49			Average Lifetime	=AVERAGE(A2:A201)	
3	19.08			Median Lifetime	=MEDIAN(A2:A201)	
4	19.11					
5	19.55					
6	18.81					
196	19.02					
197	19.72					
198	18.41					
199	17.97					
200	18.9					
201	19.54					
202						

	A	B	C	D	E	F
1	**Battery Life (hours)**					
2	19.49			Average Lifetime	18.84	
3	19.08			Median Lifetime	18.84	
4	19.11					
5	19.55					
6	18.81					
196	19.02					
197	19.72					
198	18.41					
199	17.97					
200	18.90					
201	19.54					
202						

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 insights 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 credit-card readers, bar code scanners, point-of-sale terminals, and online data collection, 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 of 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 four v's of data: volume, velocity, variety, and veracity. *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; *variety* refers to the different data types; and *veracity* refers to the reliability of the data generated.

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. However, open-source software such as R and Python are also very popular tools for performing data mining. 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 the media, 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

[2]American Statistical Association, "Ethical Guidelines for Statistical Practice," April 2018.

52 guidelines organized into eight topic areas: Professional Integrity and Accountability; Integrity of Data and Methods; Responsibilities to Science/Public/Funder/Client; Responsibilities to Research Subjects; Responsibilities to Research Team Colleagues; Responsibilities to Other Statisticians or Statistics 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 Professional Integrity and Accountability 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 Rogers Industries involving a sample of 200 lithium batteries manufactured with a new solid-state technology. The average battery life for the sample, 18.84 hours, provided an estimate of the average lifetime for all lithium batteries produced with the new solid-state technology. However, since Rogers selected a sample of batteries, it is reasonable to assume that another sample would have provided a different average battery life.

Suppose Rogers's management had hoped the sample results would enable them to claim that the average time until recharge for the new batteries was 20 hours or more. Suppose further that Rogers's management decides to continue the study by manufacturing and testing repeated samples of 200 batteries with the new solid-state technology until a sample mean of 20 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 Rogers 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 Integrity of Data and Methods area deal with issues involving the handling of data. For instance, a statistician must account for all data considered in a study and explain the sample(s) actually used. In the Rogers Industries study the average battery life for the 200 batteries in the original sample is 18.84 hours; this is less than the 20 hours or more that management hoped to obtain. Suppose now that after reviewing the results showing a 18.84 hour average battery life, Rogers discards all the observations with 18 or less hours until recharge, allegedly because these batteries contain imperfections caused by startup problems in the manufacturing process. After discarding these batteries, the average lifetime for the remaining batteries in the sample turns out to be 22 hours. Would you be suspicious of Rogers's claim that the battery life for its new solid-state batteries is 22 hours?

If the Rogers batteries showing 18 or less hours until recharge were discarded to simply provide an average lifetime of 22 hours, there is no question that discarding the batteries with 18 or fewer hours until recharge is unethical. But, even if the discarded batteries 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 Professional Integrity and Accountability 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 only sampling persons eating in restaurants that allow smoking has biased the results. If only the final results of such a study are reported, readers