END USER LICENSE AGREEMENT

This is the John Wiley and Sons, Inc. (Wiley) limited End User License Agreement, which governs your use of Wiley proprietary educational materials ("Licensed Material").

Your use of the Licensed Materials indicates your acceptance of the terms and conditions of this Agreement.

License:
Wiley hereby grants you, and you accept, a non-exclusive and non-transferable license, to use the Licensed Materials on the following terms and conditions only:

a. You acknowledge that the materials are being licensed to you for a limited term and use is subject to the terms and conditions of this Agreement. The materials have been printed pursuant to a limited Print on Demand License.

b. The Licensed Materials are for your personal use only.

c. As the Registered User, you may use the Licensed Materials for "one concurrent user" only.

d. You may not upload, copy, modify, print, transmit, transfer, sell, scan, duplicate , make or otherwise distribute or transfer the Licensed Materials or any copy or modification thereof , in whole or in part, except as expressly permitted by this Agreement. If you transfer possession of the Licensed Material or any copy or modification thereof to any third party, your license is automatically terminated. Such termination will be in addition to and not in lieu of any equitable, civil, or other remedies available to Wiley.

Term:
This License Agreement is effective for the term authorized by Wiley as specified in the Digital Content License Agreement between Wiley and your institute of Higher Education. This Agreement will also terminate upon the conditions discussed elsewhere in this Agreement, or if you fail to comply with any term or condition of this Agreement. Upon such termination, you agree to discontinue all access to the Licensed Material.

Wiley's Rights:
You acknowledge that all rights (including without limitation, copyrights, patents and trade secrets) in the Licensed Materials are the sole and exclusive property of Wiley and/or its licensors. By accepting this Agreement, you do not become the owner of the Licensed Materials, but you do have a Limited License to use them in accordance with the provisions of this Agreement. You agree to protect the Licensed Material from unauthorized use, upload, download, reproduction, or distribution. You further agree not to translate, decompile, disassemble or otherwise reverse engineer the Licensed Material.

DISCLAIMER: LICENSED MATERIALS ARE PROVIDED "AS IS," WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE RESULTS OR PERFORMANCE, AND THE COST OF ALL NECESSARY SERVICING, REPAIR, OR CORRECTION OF THE LICENSED PROGRAM AND USER MANUAL(S) IS ASSUMED BY YOU. IN NO EVENT WILL WILEY, ITS LICENSORS OR THE AUTHOR, BE LIABLE TO YOU FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION LOST PROFITS, LOST SAVINGS, OR OTHER INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE LICENSED PROGRAM OR USER MANUAL(S), EVEN IF WILEY OR AN AUTHORIZED WILEY DEALER HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

General:
This Agreement represents the entire agreement between us and supersedes any proposals or prior Agreements, oral or written, and any other communication between us relating to the subject matter of this Agreement. This Agreement will be governed and construed as if wholly entered into and performed within the State of New York. You acknowledge that you have read this Agreement, and agree to be bound by its terms and conditions.

This print copy is licensed to you pursuant to your etext license and may not be sold, copied, scanned, or duplicated. Licensee End User agrees that upon expiration or termination of this license, Licensee End User shall destroy any materials printed pursuant to this license.

Business Statistics

For Contemporary Decision Making

Ninth Edition

KEN BLACK
University of Houston—Clear Lake

WILEY

VICE PRESIDENT AND DIRECTOR	George Hoffman
EDITORIAL DIRECTOR	Veronica Visentin
EXECUTIVE EDITOR	Darren LaLonde
SPONSORING EDITOR	Jennifer Manias
EDITORIAL ASSISTANT	Ethan Lipson
PRODUCT DESIGN ASSOCIATE	Wendy Ashenberg
SENIOR CONTENT MANAGER	Dorothy Sinclair
SENIOR PRODUCTION EDITOR	Jane Lee Kaddu
EXECUTIVE MARKETING MANAGER	Christopher DeJohn
SENIOR DESIGNER	Wendy Lai
SENIOR PHOTO EDITOR	Billy Ray
COVER DESIGNER	Wendy Lai
PRODUCTION MANAGEMENT SERVICES	Aptara®
COVER IMAGE	shuoshu/iStockphoto

This book was set in 10/12 Stix by Aptara®, and printed and bound by Courier/Kendallville. The cover was printed by Courier/Kendallville.

This book is printed on acid free paper. ∞

ISBN-13: 978-1-119-32089-0

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Preface

The ninth edition of *Business Statistics for Contemporary Decision Making* continues the tradition of presenting and explaining the wonders of business statistics through the use of clear, complete, student-friendly pedagogy. The author and Wiley have vast ancillary resources available through WileyPLUS with which to complement the text in helping instructors effectively deliver this subject matter and in assisting students in their learning. With WileyPLUS instructors have far greater latitude in developing and delivering their course than ever before.

In the ninth edition, all the features of the eighth edition have been updated and changed as needed to reflect today's business world. Two new Decision Dilemmas located in Chapters 6 and 7 have been added to the ninth edition. Chapter 6, which introduces the student to continuous distributions including the normal curve, opens with a new Decision Dilemma titled "CSX Corporation," a leading rail transportation company in the U.S. Chapter 7 presents a new Decision Dilemma, "Toro" which discusses the Toro Company's 100 years of providing business solutions for the outdoor environment.

The ninth edition introduces five new cases. Chapter 2 contains the Southwest Airlines case. Chapter 6 presents a case on USAA, the United Services Automobile Association, which offers banking, investment, and insurance to people and families that serve, or served, in the United States military. Chapter 7 contains a new case on the 3M Company, best known for its adhesive products. Chapter 10 has a new case, Five Guys, a family hamburger restaurant chain with U.S. sales of $1.21 billion. Chapter 12 contains a new case on Caterpillar, Inc., the world's leading manufacturer of construction and mining equipment, diesel and natural gas engines, industrial gas turbines and diesel-electric locomotives.

All other cases in the ninth edition have been updated and edited for today's market. Three cases have been significantly revised. These include: "Statistics Describe the State of Business in India's Countryside" in Chapter 1, "Coca Cola Develops the African Market" in Chapter 3, and "Virginia Semiconductor" in Chapter 14.

Every chapter in the ninth edition contains one or two *Thinking Critically About Statistics in Business Today* features that give real-life examples of how the statistics presented in the chapter apply in the business world today. Each of these contains thought-provoking questions called *Things to Ponder* in which the objective is to effect critical thinking on behalf of the student and generate discussion involving critical thinking in the classroom. Some of these include: "Plastic Bags vs. Bringing Your Own in Japan," "Where Are Soft Drinks Sold?," "Recycling Statistics," "Probabilities in the Dry Cleaning Business," "Warehousing," and "Teleworking Facts."

This edition is designed and written for a two-semester introductory undergraduate business statistics course or an MBA-level introductory course. In addition, with 19 chapters, the ninth edition lends itself nicely to adaptation for a one-semester introductory business statistics course. The text is written with the assumption that the student has a college algebra mathematical background. No calculus is used in the presentation of material in the text.

An underlying philosophical approach to the text is that every statistical tool presented in the book has some business application. While the text contains statistical rigor, it is written so that the student can readily see that the proper application of statistics in the business world goes hand-in-hand with good decision making. In this edition, statistics are presented as a means for converting data into useful information that can be used to assist the business decision maker in making more thoughtful, information-based decisions. Thus, the text presents business statistics as "value-added" tools in the process of converting data into useful information.

Changes for the Ninth Edition

Chapters

In this edition, several changes have been made in an effort to improve the delivery and completeness of the text. The chapter organization is unchanged; however, the unit organization has been dropped. The text chapter organization allows for both one- and two-semester coverage. The first twelve chapters include the typical chapters covered in a one-semester course. The last seven chapters offer extended coverage allowing students to explore such topics as multiple regression, forecasting, analysis of categorical data, nonparametric statistics, statistical quality control, and decision analysis. For clarity and topic flow, some chapter topics have been combined and reduced in number. In Chapter 1, Statistics in Business has been included in the chapter introduction; and Variables and Data have been combined with Basic Statistical Concepts thereby reducing the number of chapter topics from four to two. In Chapter 4, Methods of Assigning Probabilities has been subsumed with Introduction to Probability.

Topical Changes

Sections and topics from the eighth edition remain virtually unchanged in the ninth edition with the exception that the former end of chapter feature, *Using the Computer*, has been removed from the text. Directions on how to use Excel and/or Minitab to compute statistics featured in the text are now housed exclusively in WileyPLUS.

In addition, in Chapter 14, the section on Logistic Regression has been significantly rewritten to reflect the changes in Minitab moving into version 17. Since much of the material on Logistic Regression is related to computer output, it was necessary to realign section 14.5 with current Minitab output for Logistic Regression. However, the conceptual explanations and the examples remain the same.

Decision Dilemma and the Decision Dilemma Solved

Each chapter of the ninth edition begins with a Decision Dilemma. Decision Dilemmas are real business vignettes that set the tone for each chapter by presenting a business dilemma and asking a number of managerial or statistical questions, the solutions to which require the use of techniques presented in the chapter. At the end of each chapter, a Decision Dilemma Solved feature discusses and answers the managerial and statistical questions posed in the Decision Dilemma using techniques from the chapter, thus bringing closure to the chapter. In the ninth edition, all decision dilemmas have been revised and updated. Solutions given in the Decision Dilemma Solved features have been revised for new data and for new versions of computer output.

In addition, two new Decision Dilemmas located in Chapters 6 and 7 have been added to the ninth edition. Chapter 6, which introduces the student to continuous distributions including the normal curve, opens with a new Decision Dilemma titled "CSX Corporation" This Decision Dilemma introduces the student to the CSX Corporation which is a leading rail transportation company in the U.S. Based in Jacksonville, Florida, the CSX Transportation network has 21,000 miles of track across the eastern half of the U.S. and two Canadian provinces. Analysis of this vignette focuses on average rail freight line-haul speed, terminal dwell time, and freight train arrival time. Students have to analyze data about these variables using probability distributions presented in Chapter 6.

Chapter 7, Sampling and Sampling Distributions, contains another new Decision Dilemma, "Toro", which discusses the Toro Company's 100 years of providing business solutions for the outdoor environment including turf, snow, and ground-engaging equipment along with irrigation and outdoor lighting solutions. The associated managerial and statistical questions tackle issues related to consumer and household spending on landscape services both in terms of dollars spent and population percentages.

Thinking Critically About Statistics in Business Today

Every chapter in the ninth edition contains at least one or two *Thinking Critically About Statistics in Business Today* features that give real-life examples of how the statistics presented in the chapter apply in the business world today. Each of these contains thought-provoking questions called Things to Ponder in which the objective is to effect critical thinking on behalf of the student and generate discussion involving critical thinking in the classroom. This approach to learning is in sync with various accreditation organizations and their current emphasis on developing critical thinking in our students. Some of these include: "Recycling Statistics," "Warehousing," "Canadian Grocery Shopping Statistics," "Beverage Consumption: America vs. Europe," "Are Facial Characteristics Correlated with CEO Traits?" "Assessing Property Values Using Multiple Regression," "Can Scrap Metal Prices Forecast the Economy?" "City Images of Cruise Destinations in the Taiwan Strait," and "Does an Iranian Auto Parts Manufacturer's Orientation Impact Innovation?" As an example, from "Beverage Consumption: America vs. Europe," Americans drink nearly five times as much soda as do Europeans and almost twice as much beer. On the other hand, Europeans drink more than twice as much tea (hot or cold), more than three times as much wine, and over four times as much tap water as Americans. Statistics show that the average American consumes forty-eight 12 oz. containers of carbonated soda per month compared to only 10 for Europeans. Europeans consume an average of sixteen 4 oz. containers of wine per month compared to an average of only five for Americans. One of the Things to Ponder questions is, "Can you think of some reasons why Americans consume more carbonated soda pop and beer than Europeans, but less wine, hot or iced tea, or tap water? Do you think that these outcomes may change in time?"

Cases

Every chapter in this text contains a unique business case. These business cases are more than just long problems; and in the discussion that follows the business scenario, several issues and questions are posed that can be addressed using techniques presented in the chapter. The ninth edition introduces five new, exciting cases. The Southwest Airlines case in chapter 2 discusses the company's rise from incorporation in 1967 to an operation with more than 3,600 flights a day, serving 94 destinations across the United States and six other countries. In Chapter 6, the reader is introduced to USAA, the United Services Automobile Association, which was founded in 1922 by 25 army officers who came together in San Antonio and decided to insure each other's automobiles. Offering banking, investment, and insurance to people and families that serve, or served, in the United States military, USAA currently has 10.7 million members, 42.6 million products, 26,300 employees and is ranked 57th in net worth ($25 billion) in Fortune 500 companies. Chapter 7 contains a new case on the 3M Company. Born as a small-scale mining company in 1902 and best known for its adhesive products, 3M is a global innovation company with over 100,000 patents, $31 billion in sales, and 90,000 employees. Five Guys, a hamburger restaurant chain whose philosophy is to serve fresh, hand-cooked burgers with "all the toppings you could stuff between fresh-baked buns" along with fresh-cut fries cooked in peanut oil, is the new case in Chapter 10. A family operation which opened its first restaurant in Arlington, Virginia in 1987 on $70,000 has grown to 1,163 U.S. restaurants with a U.S. system wide sales of $1.21 billion and units now existing in other countries. Chapter 12 contains a new case on Caterpillar, Inc. Caterpillar, headquartered in Peoria, Illinois, is an American corporation with a worldwide dealer network which sells machinery, engines, financial products and insurance. Caterpillar is the world's leading manufacturer of construction and mining equipment, diesel and natural gas engines, industrial gas turbines and diesel-electric locomotives.

All other cases in the ninth edition have been updated and edited for today's market. Three cases have been significantly revised. These include: "Statistics Describe the State of Business in India's Countryside" in Chapter 1, "Coca Cola Develops the African Market" in Chapter 3, and "Virginia Semiconductor" in Chapter 14. Here are some excerpts from one case, "The Container Store":

"In the late 1970s, Kip Tindell (chairman and CEO), Garrett Boone (Chairman Emeritus), and John Mullen (architect) drew up plans for a first-of-its-kind retail store specializing in storage solutions for both the home and the office. The vision that they created was realized when on July 1, 1978, the Container Store opened its doors in a small 1,600 square foot retail space in Dallas. The store was stocked with products that were devoted to simplifying people's lives, such as commercial parts bins, wire drawers, mailboxes, milk crates, wire leaf burners, and many others. Some critics even questioned that a store selling "empty boxes" could survive. However, the concept took off, and in the past 33 years, the company has expanded coast to coast in the United States with stores in 49 locations. Now headquartered in Coppell, Texas, the Container Store has 4,000 employees and annual revenues of over $650 million. Besides their innovative product mix, one of the keys to the success of the Container Store is the enthusiasm with which their employees work, the care that employees give to the customer, and employee knowledge of their products."

New Problems

There are 965 practice problems in the text. For ease of effort among adopters in moving from the eighth to the ninth edition, the number of problems and the problem numbers have remained the same. However, 105 of these problems have been updated to include the latest data and information available. Ten of the problems have been replaced with new problems.

Included in virtually every section of every chapter of the text is a demonstration problem which is an extra example containing both a problem and its solution and is used as an additional pedagogical tool to supplement explanations and examples in the chapters. Virtually all example and demonstration problems in the ninth edition are business oriented and contain the most current data available.

As with the previous edition, problems are located at the end of most sections in the chapters. A significant number of additional problems are provided at the end of each chapter in the Supplementary Problems. The Supplementary Problems are "scrambled"—problems using the various techniques in the chapter are mixed—so that students can test themselves on their ability to discriminate and differentiate ideas and concepts.

Databases

Available with the ninth edition are nine databases that provide additional opportunities for students to apply the statistics presented in this text. These nine databases represent a wide variety of business areas, such as agribusiness, consumer spending, energy, finance, healthcare, international labor, manufacturing, and the stock market. Altogether, these databases contain 61 variables and 7,722 observations. The data are gathered from such reliable sources as the U.S. government's Bureau of Labor, the U.S. Department of Agriculture, the American Hospital Association, the Energy Information Administration, *Moody's Handbook of Common Stocks,* and the U.S. Census Bureau. Five of the nine databases contain time-series data. The databases are 12-Year Gasoline Database, Consumer Food Database, Manufacturing Database, International Labor Database, Financial Database, Energy Database, U.S. and International Stock Market Database, Hospital Database, and Agribusiness Time-Series Database.

Video Tutorials by Ken Black

An exciting feature of the ninth edition package that will impact the effectiveness of student learning in business statistics and significantly enhance the presentation of course material is the series of videotape tutorials by Ken Black. With the advent of online business statistics courses, increasingly large class sizes, and the number of commuter students who have very limited access to educational resources on business statistics, it is often difficult for students to get the learning assistance they need to bridge the gap between theory and application on their own. There are now 21 videotaped tutorial sessions on key difficult topics in business statistics delivered by Ken Black and available for all adopters on WileyPLUS. In addition, these tutorials can easily be uploaded for classroom usage to augment lectures and enrich classroom presentations. Because there is at least one video for each of the first 12 chapters, the instructor has the option to include at least one video in the template of each chapter's plan for most, if not all, of the course. While the video tutorials vary in length, a typical video is about 10 minutes in length. The 21 video tutorials are:

1. Chapter 1: Levels of Data Measurement
2. Chapter 2: Stem-and-Leaf Plot
3. Chapter 3: Computing Variance and Standard Deviation
4. Chapter 3: Understanding and Using the Empirical Rule
5. Chapter 4: Constructing and Solving Probability Tables
6. Chapter 4: Solving Probability Word Problems
7. Chapter 5: Solving Binomial Distribution Problems, Part I
8. Chapter 5: Solving Binomial Distribution Problems, Part II
9. Chapter 6: Solving Problems Using the Normal Curve

10. Chapter 7: Solving for Probabilities of Sample Means Using the z Statistic
11. Chapter 8: Confidence Intervals
12. Chapter 8: Determining Which Inferential Technique to Use: Confidence Intervals
13. Chapter 9: Hypothesis Testing Using the z Statistic
14. Chapter 9: Establishing Hypotheses
15. Chapter 9: Understanding p-Values
16. Chapter 9: Type I and Type II errors
17. Chapter 9: Two-Tailed Tests
18. Chapter 10: Hypothesis Tests of the Difference in Means of Two Independent Populations Using the t Statistic
19. Chapter 11: Computing and Interpreting a One-Way ANOVA
20. Chapter 12: Testing the Regression Model I—Predicted Values, Residuals, and Sum of Squares of Error
21. Chapter 12: Testing the Regression Model II—Standard Error of the Estimate and r^2

Features and Benefits

Each chapter of the ninth edition contains sections called Learning Objectives, a Decision Dilemma, Demonstration Problems, Section Problems, Thinking Critically About Statistics in Business Today, Decision Dilemma Solved, Chapter Summary, Key Terms, Formulas, Ethical Considerations, Supplementary Problems, Analyzing the Databases, and Case.

- **Learning Objectives.** Each chapter begins with a statement of the chapter's main learning objectives. This statement gives the reader a list of key topics that will be discussed and the goals to be achieved from studying the chapter.
- **Decision Dilemma.** At the beginning of each chapter, a short case describes a real company or business situation in which managerial and statistical questions are raised. In most Decision Dilemmas, actual data are given and the student is asked to consider how the data can be analyzed to answer the questions.
- **Demonstration Problems.** Virtually every section of every chapter in the ninth edition contains demonstration problems. A demonstration problem contains both an example problem and its solution, and is used as an additional pedagogical tool to supplement explanations and examples.
- **Section Problems.** There are 965 problems in the text. Problems for practice are found at the end of almost every section of the text. Most problems utilize real data gathered from a plethora of sources. Included here are a few brief excerpts from some of the real-life problems in the text: "*The Wall Street Journal* reported that 40% of all workers say they would change jobs for 'slightly higher pay.' In addition, 88% of companies say that there is a shortage of qualified job candidates." "In a study by Peter D. Hart Research Associates for the Nasdaq Stock Market, it was determined that 20% of all stock investors are retired people. In addition, 40% of all U.S. adults have invested in mutual funds." "A survey conducted for the Northwestern National Life Insurance Company revealed that 70% of American workers say job stress caused frequent health problems." "According to Padgett Business Services, 20% of all small-business owners say the most important advice for starting a business is to prepare for long hours and hard work. "According to Nielsen Media Research, approximately 86% of all U.S. households have High-definition television (HDTV). In addition, 49% of all U.S. households own Digital Video Recorders (DVR)."
- **Thinking Critically About Statistics in Business Today.** Every chapter in the ninth edition contains at least one Thinking Critically About Statistics in Business Today feature. These focus boxes contain an interesting application of how techniques of that particular chapter are used in the business world today and ask probing questions of the student. They are usually based on real companies, surveys, or published research.
- **Decision Dilemma Solved.** Situated at the end of the chapter, the Decision Dilemma Solved feature addresses the managerial and statistical questions raised in the Decision Dilemma. Data given in the Decision Dilemma are analyzed computationally and by computer using techniques presented in the chapter. Answers to the managerial and statistical questions raised in the Decision Dilemma are arrived at by applying chapter concepts, thus bringing closure to the chapter.
- **Chapter Summary.** Each chapter concludes with a summary of the important concepts, ideas, and techniques of the chapter. This feature can serve as a preview of the chapter as well as a chapter review.
- **Key Terms.** Important terms are bolded and their definitions italicized throughout the text as they are discussed. At the end of the chapter, a list of the key terms from the chapter is presented. In addition, these terms appear with their definitions in the end-of-book glossary.
- **Formulas.** Important formulas in the text are highlighted to make it easy for a reader to locate them. At the end of the chapter, most of the chapter's formulas are listed together as a handy reference.
- **Ethical Considerations.** Each chapter contains an Ethical Considerations feature that is very timely, given the serious breach of ethics and lack of moral leadership of some business executives in recent years. With the abundance of statistical data and analysis, there is considerable potential for the misuse of statistics in business dealings. The important Ethical Considerations feature underscores this potential misuse by discussing such topics as lying with statistics, failing to meet statistical assumptions, and failing to include pertinent information for decision makers. Through this feature, instructors can begin to integrate the topic of ethics with applications of business statistics. Here are a few excerpts from Ethical Considerations features: "It is unprofessional and unethical to draw cause-and-effect conclusions just because two variables are correlated." "The

business researcher needs to conduct the experiment in an environment such that as many concomitant variables are controlled as possible. To the extent that this is not done, the researcher has an ethical responsibility to report that fact in the findings." "The reader is warned that the value lambda is assumed to be constant in a Poisson distribution experiment. Business researchers may produce spurious results if the value of lambda is used throughout a study, but because the study is conducted during different time periods, the value of lambda is actually changing." "In describing a body of data to an audience, it is best to use whatever statistical measures it takes to present a 'full' picture of the data. By limiting the descriptive measures used, the business researcher may give the audience only part of the picture and skew the way the receiver understands the data."

- **Supplementary Problems.** At the end of each chapter is an extensive set of additional problems. The Supplementary Problems are divided into three groups: Calculating the Statistics, which are strictly computational problems; Testing Your Understanding, which are problems for application and understanding; and Interpreting the Output, which are problems that require the interpretation and analysis of software output.
- **Analyzing the Databases.** There are nine major databases located on the student companion Web site that accompanies the ninth edition and in WileyPLUS. The end-of-chapter Analyzing the Databases section contains several questions/problems that require the application of techniques from the chapter to data in the variables of the databases. It is assumed that most of these questions/problems will be solved using a computer.
- **Case.** Each chapter has an end-of-chapter case based on a real company. These cases give the student an opportunity to use statistical concepts and techniques presented in the chapter to solve a business dilemma. Some cases feature very large companies—such as Shell Oil, Coca-Cola, or Colgate Palmolive. Others pertain to smaller businesses—such as Virginia Semiconductor, The Clarkson Company, or DeBourgh—that have overcome obstacles to survive and thrive. Most cases include raw data for analysis and questions that encourage the student to use several of the techniques presented in the chapter. In many cases, the student must analyze software output in order to reach conclusions or make decisions.

WileyPLUS

WileyPLUS with ORION is a research-based, online environment for effective teaching and learning. *WileyPLUS* builds students' confidence because it takes the guesswork out of studying by providing students with a clear roadmap: what to do, how to do it, if they did it right. This interactive approach focuses on:

Design Research-based design is based on proven instructional methods. Content is organized into small, more accessible amounts of information, helping students build better time management skills.

Engagement Students can visually track their progress as they move through the material at a pace that is right for them. Engaging in individualized self-quizzes followed by immediate feedback helps to sustain their motivation to learn.

Outcomes Self-assessment lets students know the exact outcome of their effort at any time. Advanced reporting allows instructors to easily spot trends in the usage and performance data of their class in order to make more informed decisions.

With *WileyPLUS*, students will always know:

- What to do: Features, such as the course calendar, help students stay on track and manage their time more effectively.
- How to do it: Instant feedback and personalized learning plans are available 24/7.
- If they're doing it right: Self-evaluation tools take the guesswork out of studying and help students focus on the right materials.

***WileyPLUS* for *Business Statistics, Ninth Edition* includes numerous valuable resources, among them:**

- **Ebook.** The complete text is available on WileyPLUS with learning links to various features and tools to assist students in their learning.
- **Videos.** There are 21 videos of the author explaining concepts and demonstrating how to work problems for some of the more difficult topics.
- **Applets.** Statistical applets are available, affording students the opportunity to learn concepts by iteratively experimenting with various values of statistics and parameters and observing the outcomes.
- **Learning Activities.** There are numerous learning activities to help the student better understand concepts and key terms. These activities have been developed to make learning fun, enjoyable, and challenging.
- **Data Sets.** Virtually all problems in the text along with the case problems and the databases are available to students in both Excel and Minitab format.
- **Animations.** To aid students in understanding complex interactions, selected figures from the text that involve dynamic activity have been animated. Students can download these animated figures and run them to improve their understanding of dynamic processes.
- **Flash Cards.** Key terms will be available to students in flash card format along with their definition.
- **Student Study Manual.** Complete solutions to all odd-numbered questions.
- **Demo Problems.** Step-by-step solved problems for each chapter.
- **ORION.** This adaptive, personalized learning experience delivers easy-to-use analytics so you can see exactly where your students excel and where they need help.
  - **Diagnose Early.** Simply by assigning ORION, you can diagnose the real-time proficiency of each student and see the areas that need reinforcement.

- **Facilitate Engagement.** With ORION's adaptive practice, students can interact with each other as they think more deeply about concepts at hand.
- **Measure Outcomes.** ORION helps you measure students' engagement and proficiency throughout the course so that you can easily assess how things are going at any point in time.

Ancillary Teaching and Learning Materials

www.wiley.com/college/black

Students' Companion Site

The student companion Web site contains:

- All databases in both Excel and Minitab formats for easy access and use.
- Excel and Minitab files of data from all text problems and all cases. Instructors and students now have the option of analyzing any of the data sets using the computer.
- A section on Advanced Exponential Smoothing Techniques (from Chapter 15) that offers the instructor an opportunity to delve deeper into exponential smoothing if so desired, and derivation of the slope and intercept formulas from Chapter 12.
- A tutorial on summation theory.

Instructor's Resource Kit

All instructor ancillaries are provided on the Instructor Resource Site. Included in this convenient format are:

- **Instructor's Manual.** Prepared by Ken Black, this manual contains the worked-out solutions to virtually all problems in the text. In addition, this manual contains chapter objectives, chapter outlines, chapter teaching strategies, and solutions to the cases.
- **PowerPoint Presentation Slides.** The presentation slides, contain graphics to help instructors create stimulating lectures. The PowerPoint slides may be adapted using PowerPoint software to facilitate classroom use.
- **Test Bank.** The Test Bank includes multiple-choice questions for each chapter. The Test Bank is provided in Microsoft Word format as well as in a Computerized Test Bank.

Acknowledgments

John Wiley & Sons and I would like to thank the reviewers and advisors who cared enough and took the time to provide us with their excellent insights and advice, which was used to reshape and mold the text into the ninth edition.

As always, I wish to recognize my colleagues at the University of Houston–Clear Lake for their continued interest and support of this project. In particular, I want to thank William Staples, president; Carl Stockton, provost; and Ted Cummings, dean of the School of Business for their personal interest in the book and their administrative support. A special thanks goes to Brent Goucher, my videographer, for his guidance, artistic direction, and patience in filming the videos.

There are several people within the John Wiley & Sons publishing group whom I would like to thank for their invaluable assistance on this project. These include: Lise Johnson, Darren LaLonde, Jennifer Manias, Wendy Ashenberg, Ethan Lipson, Jane Lee Kaddu, and Wendy Lai.

I want to express a special appreciation to my wife of 48 years, Carolyn, who is the love of my life and continues to provide both professional and personal support in my writing. Thanks also to my daughters, Wendi and Caycee, for their patience, love, and support.

—KEN BLACK

About the Author

KEN BLACK is currently professor of decision sciences in the School of Business at the University of Houston–Clear Lake. Born in Cambridge, Massachusetts, and raised in Missouri, he earned a bachelor's degree in mathematics from Graceland University, a master's degree in math education from the University of Texas at El Paso, a Ph.D. in business administration (management science), and a Ph.D. in educational research from the University of North Texas.

Since joining the faculty of UHCL in 1979, Professor Black has taught all levels of statistics courses, forecasting, management science, market research, and production/operations management. In 2005, he was awarded the President's Distinguished Teaching Award for the university. He has published over 20 journal articles and 20 professional papers, as well as two textbooks: *Business Statistics: An Introductory Course* and *Business Statistics for Contemporary Decision Making.* Black has consulted for many different companies, including Aetna, the city of Houston, NYLCare, AT&T, Johnson Space Center, Southwest Information Resources, UTMB, and Doctors Hospital at Renaissance. Black is active in the quality movement and is a certified Master Black Belt in Lean Six Sigma.

Ken Black and his wife, Carolyn, have two daughters, Caycee and Wendi. His hobbies include playing the guitar, reading, and traveling.

Brief Contents

CONTENTS

4 Probability 89

5 Discrete Distributions 129

6 Continuous Distributions 166

12 Simple Regression Analysis and Correlation 424

13 Multiple Regression Analysis 473

14 Building Multiple Regression Models 501

CHAPTER 1

Introduction to Statistics

LEARNING OBJECTIVES

The primary objective of Chapter 1 is to introduce you to the world of statistics, thereby enabling you to:

1. List quantitative and graphical examples of statistics within a business context.
2. Define important statistical terms, including population, sample, and parameter, as they relate to descriptive and inferential statistics.
3. Explain the difference between variables, measurement, and data.
4. Compare the four different levels of data: nominal, ordinal, interval, and ratio.

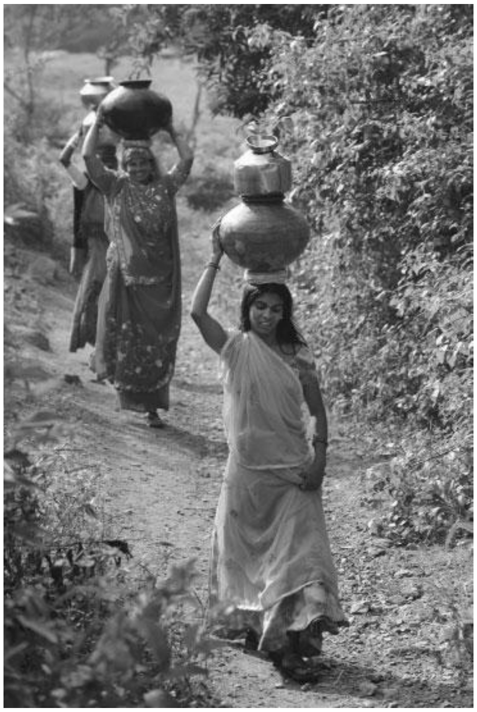
Decision Dilemma

Statistics Describe the State of Business in India's Countryside

India is the second largest country in the world with more than 1.25 billion people. More than 70% of the people live in rural areas scattered about the countryside in 650,000 villages. In fact, it can be said that one in every 10 people in the world live in rural India. While it has a per capita income of less than $1 (U.S.) per day, rural India, which has been described in the past as poor and semi-illiterate, now contributes to about one-half of the country's gross national product (GNP). However, rural India still has the most households in the world without electricity, over 300,000.

Despite its poverty and economic disadvantages, there are compelling reasons for companies to market their goods and services to rural India. The market of rural India has been growing at five times the rate of the urban India market. There is increasing agricultural productivity leading to growth in disposable income, and there is a reduction in the gap between the tastes of urban and rural customers. The literacy level is increasing, and people are becoming more conscious about their lifestyles and opportunities for a better life.

Around 60% of all middle-income households in India are in rural areas and more than one-third of all rural households in India now have a main source of income other than farming. Virtually every home has a radio, about one-third have a television, and more than one-half of rural households benefit from banking services. Forty-two percent of the people living in India's villages and small towns use toothpastes and that proportion is increasing as rural income rises and as there is greater awareness about oral hygiene.



© Hemis/Alamy Stock Photo

In rural India, consumers are gaining more disposable income due to the movement of manufacturing jobs to rural areas. It is estimated that nearly 75% of the factories that opened in India in the past decade were built in rural areas. Products that are doing well in sales to people in rural India include televisions, fans, bicycles, bath soap, two- or three-wheelers, cars, and many others. According to MART, a New Delhi-based research organization, rural India buys 46% of all soft drinks and 49% of motorcycles sold in India. Because of such factors, many U.S. and Indian firms such as Microsoft, General Electric, Kellogg's, Colgate-Palmolive, Idea Cellular, Hindustan Lever, Godrej, Nirma Chemical Works, Novartis, Dabur, Tata Motors, and Vodafone India have entered the rural Indian market with enthusiasm. Marketing to rural customers often involves persuading them to try and to adopt products that they may not have used before. Rural India is a huge, relatively untapped market for businesses. However, entering such a market is not without risks and obstacles. The dilemma facing companies is whether to enter this marketplace and, if so, to what extent and how.

MANAGERIAL AND STATISTICAL QUESTIONS

1. Are the statistics presented in this report exact figures or estimates?
2. How and where could researchers have gathered such data?
3. In measuring the potential of the rural India marketplace, what other statistics could have been gathered?
4. What levels of data measurement are represented by data on rural India?
5. How can managers use these and other statistics to make better decisions about entering this marketplace?

Source: Adapted from "Marketing to Rural India: Making the Ends Meet," March 8, 2007, in *India Knowledge@Wharton*, http://knowledge.wharton.upenn.edu/india/article.cfm?articleid=4172; "Rural Segment Quickly Catching Up", September 2015, IBEF (India Brand Equity Foundation) at: www.ibef.org/industry/indian-rural-market.aspx; "Unlocking the Wealth in Rural Markets", June 2014, *Harvard Business Review* at: https://hbr.org/2014/06/unlocking-the-wealth-in-rural-markets; "Much of Rural India Still Waits for Electricity", October 2013, University of Washington, at: artsci.washington.edu/news/2013-10/much-rural-india-still-waits-electricity

Every minute of the working day, decisions are made by businesses around the world that determine whether companies will be profitable and growing or whether they will stagnate and die. Most of these decisions are made with the assistance of information gathered about the marketplace, the economic and financial environment, the workforce, the competition, and other factors. Such information usually comes in the form of data or is accompanied by data. Business statistics provides the tool through which such data are collected, analyzed, summarized, and presented to facilitate the decision-making process, and business statistics plays an important role in the ongoing saga of decision making within the dynamic world of business.

Virtually every area of business uses statistics in decision making. Here are some recent examples:

- According to a national survey of independent business owners conducted by the Institute for Local Self-Reliance in partnership with the Advocates for Independent Business coalition, when asked "Which two public policy changes would most help your business?" (retailers only), 40% said "Pass the Marketplace Fairness Act" and 38% said "Cap Credit Card Swipe Fees".
- A survey of 1465 workers by Hotjobs reports that 55% of workers believe that the quality of their work is perceived the same when they work remotely as when they are physically in the office.
- A survey of 477 executives by the Association of Executive Search Consultants determined that 48% of men and 67% of women say they are more likely to negotiate for less business travel compared with five years ago.
- A global Family Business Survey of 2,378 respondents sponsored by PwC reported that 65% of family businesses reported growth in the last twelve months and 49% of respondents are apprehensive about their ability to recruit skilled staff in the next twelve months.
- A Deloitte Retail "Green" survey of 1080 adults revealed that 54% agreed that plastic, non-compostable shopping bags should be banned.
- A study of consumer electronics spending by a 2,500 member on-line panel of the NPD group showed that consumers expect to spend $555, on average, per person on new consumer electronics devices this year.

You can see from these few examples that there is a wide variety of uses and applications of statistics in business. Note that in most of these examples, business researchers have conducted a study and provided us rich and interesting information.

Nick M Do/Photodisc/ Getty Images, Inc.

William King/The Image Bank/ Getty Images, Inc.

Comstock/Getty Images, Inc.

In this text we will examine several types of graphs for depicting data as we study ways to arrange or structure data into forms that are both meaningful and useful to decision makers. We will learn about techniques for sampling from a population that allow studies of the business world to be conducted more inexpensively and in a more timely manner. We will explore various ways to forecast future values and examine techniques for predicting trends. This text also includes many statistical tools for testing hypotheses and for estimating population values. These and many other exciting statistics and statistical techniques await us on this journey through business statistics. Let us begin.

1.1 Basic Statistical Concepts

Business statistics, like many areas of study, has its own language. It is important to begin our study with an introduction of some basic concepts in order to understand and communicate about the subject. We begin with a discussion of the word *statistics*. The word *statistics* has many different meanings in our culture. *Webster's Third New International Dictionary* gives a comprehensive definition of **statistics** as *a science dealing with the collection, analysis, interpretation, and presentation of numerical data*. Viewed from this perspective, statistics includes all the topics presented in this text. **Figure 1.1** graphically displays the key elements of statistics.

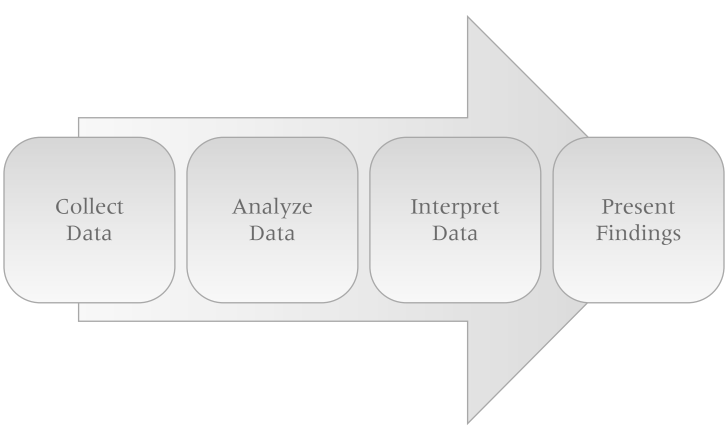


FIGURE 1.1 **The Key Elements of Statistics**

The study of statistics can be organized in a variety of ways. One of the main ways is to subdivide statistics into two branches: descriptive statistics and inferential statistics. To understand the difference between descriptive and inferential statistics, definitions of *population* and *sample* are helpful. *Webster's Third New International Dictionary* defines **population** as *a collection of persons, objects, or items of interest*. The population can be a widely defined category, such as "all automobiles," or it can be narrowly defined, such as "all Ford Mustang cars produced from 2014 to 2016." A population can be a group of people, such as "all workers presently employed by Microsoft," or it can be a set of objects, such as "all dishwashers produced on February 3, 2016, by the General Electric Company at the Louisville plant." The researcher defines the population to be whatever he or she is studying. When researchers *gather data from the whole population for a given measurement of interest,* they call it a **census**. Most people are familiar with the U.S. Census. Every 10 years, the government attempts to measure all persons living in this country.

A **sample** is *a portion of the whole* and, if properly taken, is representative of the whole. For various reasons (explained in Chapter 7), researchers often prefer to work with a sample of the population instead of the entire population. For example, in conducting quality-control experiments to determine the average life of lightbulbs, a lightbulb manufacturer might randomly sample only 75 lightbulbs during a production run. Because of time and money limitations, a human resources manager might take a random sample of 40 employees instead of using a census to measure company morale.

If a business analyst is *using data gathered on a group to describe or reach conclusions about that same group,* the statistics are called **descriptive statistics**. For example, if an instructor produces statistics to summarize a class's examination effort and uses those statistics to reach conclusions about that class only, the statistics are descriptive.

Many of the statistical data generated by businesses are descriptive. They might include number of employees on vacation during June, average salary at the Denver office, corporate sales for 2016, average managerial satisfaction score on a company-wide census of employee attitudes, and average return on investment for the Lofton Company for the years 1996 through 2016.

Another type of statistics is called **inferential statistics**. If a researcher *gathers data from a sample and uses the statistics generated to reach conclusions about the population from which the sample was taken,* the statistics are inferential statistics. The data gathered from the sample are used to infer something about a larger group. Inferential statistics are sometimes referred to as *inductive statistics.* The use and importance of inferential statistics continue to grow.

One application of inferential statistics is in pharmaceutical research. Some new drugs are expensive to produce, and therefore tests must be limited to small samples of patients. Utilizing inferential statistics, researchers can design experiments with small randomly selected samples of patients and attempt to reach conclusions and make inferences about the population.

Market researchers use inferential statistics to study the impact of advertising on various market segments. Suppose a soft drink company creates an advertisement depicting a dispensing machine that talks to the buyer, and market researchers want to measure the impact of the new advertisement on various age groups. The researcher could stratify the population into age categories ranging from young to old, randomly sample each stratum, and use inferential statistics to determine the effectiveness of the advertisement for the various age groups in the population. The advantage of using inferential statistics is that they enable the researcher to study effectively a wide range of phenomena without having to conduct a census. Most of the topics discussed in this text pertain to inferential statistics.

A *descriptive measure of the population* is called a **parameter**. Parameters are usually denoted by Greek letters. Examples of parameters are population mean (μ), population variance (σ^2), and population standard deviation (σ). A *descriptive measure of a sample* is called a **statistic**. Statistics are usually denoted by Roman letters. Examples of statistics are sample mean ($\bar{x}$), sample variance (s^2), and sample standard deviation (s).

Differentiation between the terms *parameter* and *statistic* is important only in the use of inferential statistics. A business researcher often wants to estimate the value of a parameter or conduct tests about the parameter. However, the calculation of parameters is usually either impossible or infeasible because of the amount of time and money required to take a census.

In such cases, the business researcher can take a random sample of the population, calculate a statistic on the sample, and infer by estimation the value of the parameter. The basis for inferential statistics, then, is the ability to make decisions about parameters without having to complete a census of the population.

For example, a manufacturer of washing machines would probably want to determine the average number of loads that a new machine can wash before it needs repairs. The parameter is the population mean or average number of washes per machine before repair. A company researcher takes a sample of machines, computes the number of washes before repair for each machine, averages the numbers, and estimates the population value or parameter by using the statistic, which in this case is the sample average. **Figure 1.2** demonstrates the inferential process.

Inferences about parameters are made under uncertainty. Unless parameters are computed directly from the population, the statistician never knows with certainty whether the estimates or inferences made from samples are true. In an effort to estimate the level of confidence in the result of the process, statisticians use probability statements. For this and other reasons, part of this text is devoted to probability (Chapter 4).

Business statistics is about measuring phenomena in the business world and organizing, analyzing, and presenting the resulting numerical information in such a way such that better, more informed business decisions can be made. Most business statistics studies contain variables, measurements, and data.

In business statistics, a **variable** is *a characteristic of any entity being studied that is capable of taking on different values*. Some examples of variables in business might include return on investment, advertising dollars, labor productivity, stock price, historical cost, total sales, market share, age of worker, earnings per share, miles driven to work, time spent in store shopping, and many, many others. In business statistics studies, most variables produce a measurement that can be used for analysis. A **measurement** is *when a standard process is used to assign numbers to particular attributes or characteristics of a variable.* Many measurements are obvious, such as time spent in a store shopping by a customer, age of the worker, or the number of miles driven to work. However, some measurements, such as labor productivity, customer satisfaction, and return on investment, have to be defined by the business researcher or by experts within the field. Once such measurements are recorded and stored, they can be denoted as "data." It can be said that **data** are *recorded measurements.* The processes of measuring and data gathering are basic to all that we do in business statistics. It is data that are analyzed by a business statistician in order to learn more about the variables being studied. Sometimes, sets of data are organized into databases as a way to store data or as a means for more conveniently analyzing data or comparing variables. Valid data

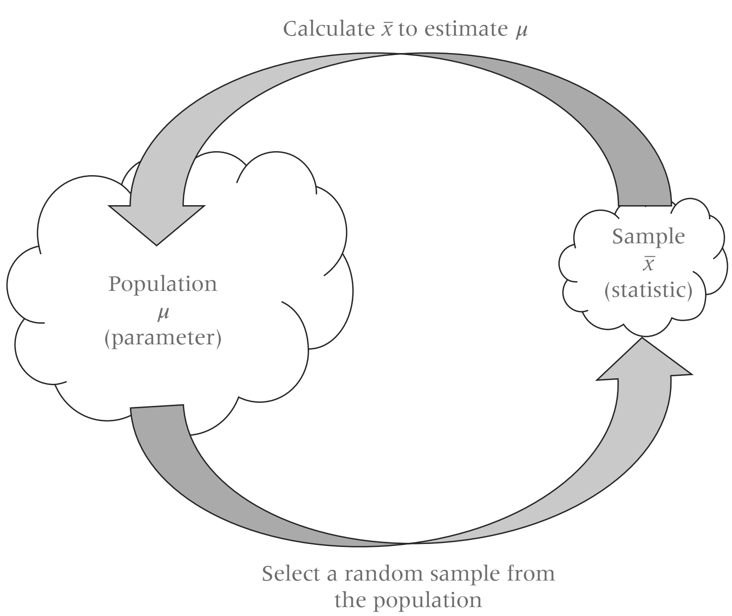


FIGURE 1.2 **The Inferential Process**

are the lifeblood of business statistics, and it is important that the business researcher give thoughtful attention to the creation of meaningful, valid data before embarking on analysis and reaching conclusions.

1.2 Data Measurement

Lecture Video

Millions of numerical data are gathered in businesses every day, representing myriad items. For example, numbers represent dollar costs of items produced, geographical locations of retail outlets, weights of shipments, and rankings of subordinates at yearly reviews. All such data should not be analyzed the same way statistically because the entities represented by the numbers are different. For this reason, the business researcher needs to know the *level of data measurement* represented by the numbers being analyzed.

The disparate use of numbers can be illustrated by the numbers 40 and 80, which could represent the weights of two objects being shipped, the ratings received on a consumer test by two different products, or football jersey numbers of a fullback and a wide receiver. Although 80 pounds is twice as much as 40 pounds, the wide receiver is probably not twice as big as the fullback! Averaging the two weights seems reasonable, but averaging the football jersey numbers makes no sense. The appropriateness of the data analysis depends on the level of measurement of the data gathered. The phenomenon represented by the numbers determines the level of data measurement. Four common levels of data measurement follow.

1. Nominal
2. Ordinal
3. Interval
4. Ratio

Nominal is the lowest level of data measurement followed by ordinal, interval, and ratio. Ratio is the highest level of data measurement, as shown in **Figure 1.3**.

Nominal Level

The *lowest level of data measurement* is the **nominal level**. Numbers representing nominal-level data (the word *level* often is omitted) can be *used only to classify or categorize.* Employee identification numbers are an example of nominal data. The numbers are used only to differentiate employees and not to make a value statement about them. Many demographic questions in surveys result in data that are nominal because the questions are used for classification only. The following is an example of such a question that would result in nominal data:

Which of the following employment classifications best describes your area of work?

1. Educator
2. Construction worker
3. Manufacturing worker
4. Lawyer
5. Doctor
6. Other

Suppose that, for computing purposes, an educator is assigned a 1, a construction worker is assigned a 2, a manufacturing worker is assigned a 3, and so on. These numbers should be used only to classify respondents. The number 1 does not denote the top classification. It is used only to differentiate an educator (1) from a lawyer (4).

Some other types of variables that often produce nominal-level data are sex, religion, ethnicity, geographic location, and place of birth. Social Security numbers, telephone

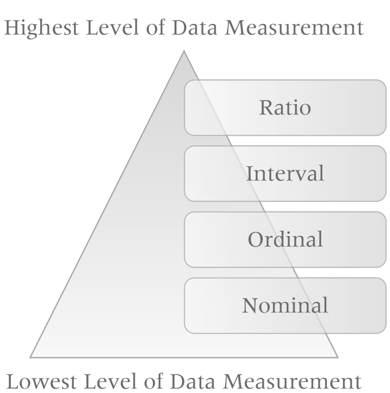


FIGURE 1.3 Hierarchy of Levels of Data

numbers, employee ID numbers, and ZIP code numbers are further examples of nominal data. Statistical techniques that are appropriate for analyzing nominal data are limited. However, some of the more widely used statistics, such as the chi-square statistic, can be applied to nominal data, often producing useful information.

Ordinal Level

Ordinal-level data measurement is higher than the nominal level. In addition to the nominal-level capabilities, ordinal-level measurement can be used to rank or order people or objects. For example, using ordinal data, a supervisor can evaluate three employees by ranking their productivity with the numbers 1 through 3. The supervisor could identify one employee as the most productive, one as the least productive, and one as somewhere between by using ordinal data. However, the supervisor could not use ordinal data to establish that the intervals between the employees ranked 1 and 2 and between the employees ranked 2 and 3 are equal; that is, she could not say that the differences in the amount of productivity between workers ranked 1, 2, and 3 are necessarily the same. With ordinal data, the distances or spacing represented by consecutive numbers are not always equal.

Some questionnaire Likert-type scales are considered by many researchers to be ordinal in level. The following is an example of one such scale:

This computer tutorial is	___	___	___	___	___
	not helpful	somewhat helpful	moderately helpful	very helpful	extremely helpful
	1	2	3	4	5

When this survey question is coded for the computer, only the numbers 1 through 5 will remain, not the adjectives. Virtually everyone would agree that a 5 is higher than a 4 on this scale and that ranking responses is possible. However, most respondents would not consider the differences between not helpful, somewhat helpful, moderately helpful, very helpful, and extremely helpful to be equal.

Mutual funds as investments are sometimes rated in terms of risk by using measures of default risk, currency risk, and interest rate risk. These three measures are applied to investments by rating them as having high, medium, and low risk. Suppose high risk is assigned a 3, medium risk a 2, and low risk a 1. If a fund is awarded a 3 rather than a 2, it carries more risk, and so on. However, the differences in risk between categories 1, 2, and 3 are not necessarily equal. Thus, these measurements of risk are only ordinal-level measurements. Another example of the use of ordinal numbers in business is the ranking of the top 50 most admired companies in *Fortune* magazine. The numbers ranking the companies are only ordinal in measurement. Certain statistical techniques are specifically suited to ordinal data, but many other techniques are not appropriate for use on ordinal data. For example, it does not make sense to say that the average of “moderately helpful” and “very helpful” is “moderately helpful and a half.”

Because nominal and ordinal data are often derived from imprecise measurements such as demographic questions, the categorization of people or objects, or the ranking of items, *nominal and ordinal data* are **nonmetric data** and are sometimes referred to as *qualitative data.*

Interval Level

Interval-level data measurement is the *next to the highest level of data in which the distances between consecutive numbers have meaning and the data are always numerical.* The distances represented by the differences between consecutive numbers are equal; that is, interval data have equal intervals. An example of interval measurement is Fahrenheit temperature. With Fahrenheit temperature numbers, the temperatures can be ranked, and the amounts of heat between consecutive readings, such as 20°, 21°, and 22°, are the same.

In addition, with interval-level data, the zero point is a matter of convention or convenience and not a natural or fixed zero point. Zero is just another point on the scale and does not mean the absence of the phenomenon. For example, zero degrees Fahrenheit is not the lowest possible temperature. Some other examples of interval-level data are the percentage change in employment, the percentage return on a stock, and the dollar change in stock price.

Ratio Level

Ratio-level data measurement is *the highest level of data measurement.* Ratio data *have the same properties as interval data,* but ratio data have an *absolute zero,* and *the ratio of two numbers is meaningful.* The notion of absolute zero means that zero is fixed, and *the zero value in the data represents the absence of the characteristic being studied.* The value of zero cannot be arbitrarily assigned because it represents a fixed point. This definition enables the statistician to create *ratios* with the data.

Examples of ratio data are height, weight, time, volume, and Kelvin temperature. With ratio data, a researcher can state that 180 pounds of weight is twice as much as 90 pounds or, in other words, make a ratio of 180:90. Many of the data measured by valves or gauges in industry are ratio data.

Other examples in the business world that are ratio level in measurement are production cycle time, work measurement time, passenger miles, number of trucks sold, complaints per 10,000 fliers, and number of employees.

Because interval- and ratio-level data are usually gathered by precise instruments often used in production and engineering processes, in national standardized testing, or in standardized accounting procedures, they are called **metric data** and are sometimes referred to as *quantitative* data.

Comparison of the Four Levels of Data

Figure 1.4 shows the relationships of the usage potential among the four levels of data measurement. The concentric squares denote that each higher level of data can be analyzed by any of the techniques used on lower levels of data but, in addition, can be used in other statistical techniques. Therefore, ratio data can be analyzed by any statistical technique applicable to the other three levels of data plus some others.

Nominal data are the most limited data in terms of the types of statistical analysis that can be used with them. Ordinal data allow the researcher to perform any analysis that can be done with nominal data and some additional analyses. With ratio data, a statistician can make ratio comparisons and appropriately do any analysis that can be performed on nominal, ordinal, or interval data. Some statistical techniques require ratio data and cannot be used to analyze other levels of data.

Statistical techniques can be separated into two categories: parametric statistics and nonparametric statistics. **Parametric statistics** require that data be interval or ratio. If the data are nominal or ordinal, **nonparametric statistics** must be used. Nonparametric statistics can also be used to analyze interval or ratio data. This text focuses largely on parametric statistics, with the exception of Chapter 16 and Chapter 17, which contain nonparametric techniques. Thus much of the material in this text requires that data be interval or ratio data.

Figure 1.5 contains a summary of metric data and nonmetric data.

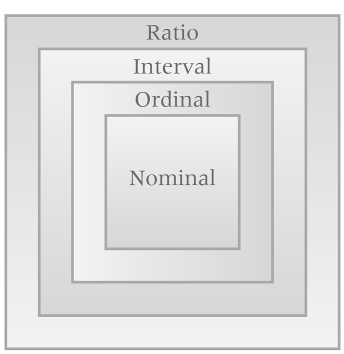


FIGURE 1.4 **Usage Potential of Various Levels of Data**

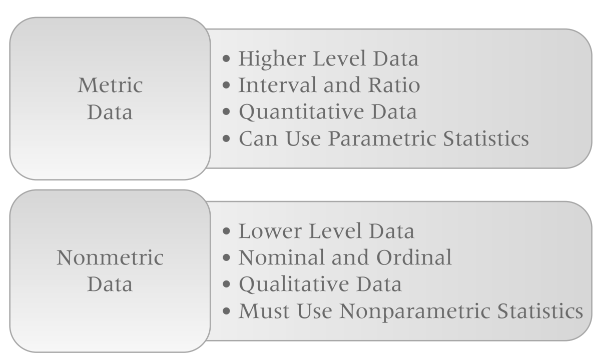


FIGURE 1.5 **Metric vs. Nonmetric Data**

DEMONSTRATION PROBLEM 1.1

Because of increased competition for patients among providers and the need to determine how providers can better serve their clientele, hospital administrators sometimes administer a quality satisfaction survey to their patients after the patient is released. The following types of questions are sometimes asked on such a survey. These questions will result in what level of data measurement?

1. How long ago were you released from the hospital?

2. Which type of unit were you in for most of your stay?

__Coronary care
__Intensive care
__Maternity care
__Medical unit
__Pediatric/children's unit
__Surgical unit

3. In choosing a hospital, how important was the hospital's location?

(circle one)

Very Important	Somewhat Important	Not Very Important	Not at All Important

4. What was your body temperature when you were admitted to the hospital?

5. Rate the skill of your doctor:

__Excellent __Very Good __Good __Fair __Poor

Solution: Question 1 is a time measurement with an absolute zero and is therefore ratio-level measurement. A person who has been out of the hospital for two weeks has been out twice as long as someone who has been out of the hospital for one week.

Question 2 yields nominal data because the patient is asked only to categorize the type of unit he or she was in. This question does not require a hierarchy or ranking of the type of unit. Questions 3 and 5 are likely to result in ordinal-level data. Suppose a number is assigned the descriptors in these two questions. For question 3, "very important" might be assigned a 4, "somewhat important" a 3, "not very important" a 2, and "not at all important" a 1. Certainly, the higher the number, the more important is the hospital's location. Thus, these responses can be ranked by selection. However, the increases in importance from 1 to 2 to 3 to 4 are not necessarily equal. This same logic applies to the numeric values assigned in question 5. In question 4, body temperature, if measured on a Fahrenheit or Celsius scale, is interval in measurement.

Statistical Analysis Using the Computer: Excel and Minitab

The advent of the modern computer opened many new opportunities for statistical analysis. The computer allows for storage, retrieval, and transfer of large data sets. Furthermore, computer software has been developed to analyze data by means of sophisticated statistical techniques. Some widely used statistical techniques, such as multiple regression, are so tedious and cumbersome to compute manually that they were of little practical use to researchers before computers were developed.

Business statisticians use many popular statistical software packages, including Minitab, SAS, and SPSS. Many computer spreadsheet software packages also have the capability of analyzing data statistically. In this text, the computer statistical output presented is from both the Minitab and the Microsoft Excel software.

Thinking Critically About Statistics in Business Today

Cellular Phone Use in Japan

The Communications and Information Network Association of Japan (CIAJ) conducts an annual study of cellular phone use in Japan. This past year a survey was taken of 1200 mobile phone users split evenly between men and women and almost equally over six age brackets of people residing in the larger Tokyo and Osaka metropolitan areas. The study produced several interesting findings. Of the respondents, 76.8% said that their main-use terminal was a smartphone while 23.2% said it was a feature phone. Of the smartphone users, 49.2% reported that their second device was a tablet (with a telecom subscription) compared to the previous year where 36.7% said that a tablet was their second device. Survey participants were asked what are the decisive factors in purchasing smart phone? The number one decisive factor was battery life, selected by over 70%, followed by manufacturer/brand selected by 67.5%. Survey participants who owned feature phones were asked the same question and the number one decisive factor was monthly payment cost (84.7%) followed by purchase price of handset (84.4%). In terms of usage of smart phone features and services, the number one most popular use was "searching the Internet" (90.2%) followed by "shooting still photos with camera" selected by 88.7%.

THINGS TO PONDER

1. In what way was this study an example of inferential statistics?
2. What is the population of this study?
3. What are some of the variables being studied?
4. How might a study such as this yield information that is useful to business decision makers?

Source: "CIAJ Releases Report on the Study of Mobile Phone Use", July 29, 2015 at http://www.ciaj.or.jp/en/news/news2015/2015/07/29/951/

Decision Dilemma Solved

Statistics Describe the State of Business in India's Countryside

Several statistics were reported in the Decision Dilemma about rural India. The authors of the sources from which the Decision Dilemma was drawn never stated whether the reported statistics were based on actual data drawn from a census of rural India households or on estimates taken from a sample of rural households. If the data came from a census, then the totals, averages, and percentages presented in the Decision Dilemma are parameters. If, on the other hand, the data were gathered from samples, then they are statistics. Although governments especially do conduct censuses and at least some of the reported numbers could be parameters, more often than not, such data are gathered from samples of people or items. For example, in rural India, the government, academicians, or business researchers could have taken random samples of households, gathering consumer statistics that are then used to estimate population parameters, such as percentage of households with televisions, and so forth.

In conducting research on a topic like consumer consumption in rural India, there is potential for a wide variety of statistics to be gathered that represent several levels of data. For example, ratio-level measurements on items such as income, number of children, age of household heads, number of livestock, and grams of toothpaste consumed per year might be obtained. On the other hand, if researchers use a Likert scale (1-to-5 measurements) to gather responses about the interests, likes, and preferences of rural India consumers, an ordinal-level measurement would be obtained, as would the ranking of products or brands in market research studies. Other variables, such as geographic location, sex, occupation, or religion, are usually measured with nominal data.

The decision to enter the rural India market is not just a marketing decision. It involves production capacity and schedule issues, transportation challenges, financial commitments, managerial growth or reassignment, accounting issues (accounting for rural India may differ from techniques used in traditional markets), information systems, and other related areas. With so much on the line, company decision makers need as much relevant information available as possible. In this Decision Dilemma, it is obvious to the decision maker that rural India is still quite poor and illiterate. Its capacity as a market is great. The statistics on the increasing sales of a few personal-care products look promising. What are the future forecasts for the earning power of people in rural India? Will major cultural issues block the adoption of the types of products that companies want to sell there? The answers to these and many other interesting and useful questions can be obtained by the appropriate use of statistics. The 800 million people living in rural India represent the second largest group of people in the world. It certainly is a market segment worth studying further.

Ethical Considerations

With the abundance and proliferation of statistical data, potential misuse of statistics in business dealings is a concern. It is, in effect, unethical business behavior to use statistics out of context. Unethical business people might use only selective data from studies to underscore their point, omitting statistics from the same studies that argue against their case. The results of statistical studies can be misstated or overstated to gain favor.

This chapter noted that if data are nominal or ordinal, then only nonparametric statistics are appropriate for analysis. The use of parametric statistics to analyze nominal and/or ordinal data is wrong and could be considered under some circumstances to be unethical.

In this text, each chapter contains a section on ethics that discusses how business people can misuse the techniques presented in the chapter in an unethical manner. As both users and producers, business students need to be aware of the potential ethical pitfalls that can occur with statistics.

Summary

Statistics is an important decision-making tool in business and is used in virtually every area of business. In this course, the word *statistics* is defined as the science of gathering, analyzing, interpreting, and presenting numerical data.

The study of statistics can be subdivided into two main areas: *descriptive statistics* and *inferential statistics.* Descriptive statistics result from gathering data from a body, group, or population and reaching conclusions only about that group. Inferential statistics are generated by gathering sample data from a group, body, or population and reaching conclusions about the larger group from which the sample was drawn.

Most business statistics studies contain variables, measurements, and data. A *variable* is a characteristic of any entity being studied that is capable of taking on different values. Examples of variables might include monthly household food spending, time between arrivals at a restaurant, and patient satisfaction rating. A *measurement* is when a standard process is used to assign numbers to particular attributes or characteristics of a variable. Measurements on monthly household food spending might be taken in dollars, time between arrivals might be measured in minutes, and patient satisfaction might be measured using a 5-point scale. *Data* are recorded measurements. It is data that are analyzed by business statisticians in order to learn more about the variables being studied.

The appropriate type of statistical analysis depends on the level of data measurement, which can be (1) *nominal,* (2) *ordinal,* (3) *interval,* or (4) *ratio.* Nominal is the lowest level, representing classification only of such data as geographic location, sex, or Social Security number. The next level is ordinal, which provides rank ordering measurements in which the intervals between consecutive numbers do not necessarily represent equal distances. Interval is the next to highest level of data measurement in which the distances represented by consecutive numbers are equal. The highest level of data measurement is ratio, which has all the qualities of interval measurement, but ratio data contain an absolute zero and ratios between numbers are meaningful. Interval and ratio data sometimes are called *metric* or *quantitative* data. Nominal and ordinal data sometimes are called *nonmetric* or *qualitative* data.

Two major types of inferential statistics are (1) *parametric statistics* and (2) *nonparametric statistics.* Use of parametric statistics requires interval or ratio data and certain assumptions about the distribution of the data. The techniques presented in this text are largely parametric. If data are only nominal or ordinal in level, nonparametric statistics must be used.

Key Terms

census
data
descriptive statistics
inferential statistics
interval-level data
measurement
metric data
nominal-level data
nonmetric data
nonparametric statistics
ordinal-level data
parameter
parametric statistics
population
ratio-level data
sample
statistic
statistics
variable

Supplementary Problems

1.1. Give a specific example of data that might be gathered from each of the following business disciplines: accounting, finance, human resources, marketing, information systems, production, and management. An example in the marketing area might be "number of sales per month by each salesperson."

1.2 State examples of data that can be gathered for decision making purposes from each of the following industries: manufacturing, insurance, travel, retailing, communications, computing, agriculture, banking, and healthcare. An example in the travel industry might be the cost of business travel per day in various European cities.

1.3 Give an example of *descriptive* statistics in the recorded music industry. Give an example of how *inferential* statistics could be used in the recorded music industry. Compare the two examples. What makes them different?

1.4 Suppose you are an operations manager for a plant that manufactures batteries. Give an example of how you could use *descriptive* statistics to make better managerial decisions. Give an example of how you could use *inferential* statistics to make better managerial decisions.

1.5 There are many types of information that might help the manager of a large department store run the business more efficiently and better understand how to improve sales. Think about this in such areas as sales, customers, human resources, inventory, suppliers, etc., and list five variables that might produce information that could aid the manager in his or her job. Write a sentence or two describing each variable, and briefly discuss some numerical observations that might be generated for each variable.

1.6 Suppose you are the owner of a medium-sized restaurant in a small city. What are some variables associated with different aspects of the business that might be helpful to you in making business decisions about the restaurant? Name four of these variables, and for each variable, briefly describe a numerical observation that might be the result of measuring the variable.

1.7 Classify each of the following as nominal, ordinal, interval, or ratio data.

- **a.** The time required to produce each tire on an assembly line
- **b.** The number of quarts of milk a family drinks in a month
- **c.** The ranking of four machines in your plant after they have been designated as excellent, good, satisfactory, and poor
- **d.** The telephone area code of clients in the United States
- **e.** The age of each of your employees
- **f.** The dollar sales at the local pizza shop each month
- **g.** An employee's identification number
- **h.** The response time of an emergency unit

1.8 Classify each of the following as nominal, ordinal, interval, or ratio data.

- **a.** The ranking of a company in the *Fortune* 500
- **b.** The number of tickets sold at a movie theater on any given night
- **c.** The identification number on a questionnaire
- **d.** Per capita income
- **e.** The trade balance in dollars
- **f.** Profit/loss in dollars
- **g.** A company's tax identification
- **h.** The Standard & Poor's bond ratings of cities based on the following scales:

RATING	GRADE
Highest quality	AAA
High quality	AA
Upper medium quality	A
Medium quality	BBB
Somewhat speculative	BB
Low quality, speculative	B
Low grade, default possible	CCC
Low grade, partial recovery possible	CC
Default, recovery unlikely	C

1.9 The Rathburn Manufacturing Company makes electric wiring, which it sells to contractors in the construction industry. Approximately 900 electric contractors purchase wire from Rathburn annually. Rathburn's director of marketing wants to determine electric contractors' satisfaction with Rathburn's wire. He developed a questionnaire that yields a satisfaction score between 10 and 50 for participant responses. A random sample of 35 of the 900 contractors is asked to complete a satisfaction survey. The satisfaction scores for the 35 participants are averaged to produce a mean satisfaction score.

- **a.** What is the population for this study?
- **b.** What is the sample for this study?
- **c.** What is the statistic for this study?
- **d.** What would be a parameter for this study?

Analyzing the Databases

See www.wiley.com/college/black

Nine databases are available with this text, providing additional opportunities to apply the statistics presented in this course. These databases are located in WileyPLUS, and each is available in either Minitab or Excel format for your convenience. These nine databases represent a wide variety of business areas, such as agribusiness, consumer spending, energy, finance, healthcare, international labor, manufacturing, and the stock market. Altogether, these databases contain 61 variables and 7722 observations. The data are gathered from such reliable sources as the U.S. government's Bureau of Labor, the U.S. Department of Agriculture, the American Hospital Association, the Energy Information Administration, *Moody's Handbook of Common Stocks,* and the U.S. Census Bureau. Five of the nine databases contain time-series data. These databases are:

12-Year Gasoline Database

The 12-year time-series gasoline database contains monthly data for four variables: U.S. Gasoline Prices, OPEC Spot Price, U.S. Finished Motor Gasoline Production, and U.S. Natural Gas Wellhead Price. There are 137 data entries for each variable. U.S. Gasoline Prices are given in cents, the OPEC Spot Price is given in dollars per barrel, U.S. Finished Motor Gasoline Production is given in 1000 barrels per day, and U.S. Natural Gas Wellhead Price is given in dollars per 1000 cubic feet.

Consumer Food Database

The consumer food database contains five variables: Annual Food Spending per Household, Annual Household Income, Non-Mortgage Household Debt, Geographic Region of the U.S. of the Household, and Household Location. There are 200 entries for each variable in this database representing 200 different households from various regions and locations in the United States. Annual Food Spending per Household, Annual Household Income, and Non-Mortgage Household Debt are all given in dollars. The variable Region tells in which one of four regions the household resides. In this variable, the Northeast is coded as 1, the Midwest is coded 2, the South is coded as 3, and the West is coded as 4. The variable Location is coded as 1 if the household is in a metropolitan area and 2 if the household is outside a metro area. The data in this database were randomly derived and developed based on actual national norms.

Manufacturing Database

This database contains eight variables taken from 20 industries and 140 subindustries in the United States. Some of the industries are food products, textile mill products, furniture, chemicals, rubber products, primary metals, industrial machinery, and transportation equipment. The eight variables are Number of Employees, Number of Production

Workers, Value Added by Manufacture, Cost of Materials, Value of Industry Shipments, New Capital Expenditures, End-of-Year Inventories, and Industry Group. Two variables, Number of Employees and Number of Production Workers, are in units of 1000. Four variables, Value Added by Manufacture, Cost of Materials, New Capital Expenditures, and End-of-Year Inventories, are in million-dollar units. The Industry Group variable consists of numbers from 1 to 20 to denote the industry group to which the particular subindustry belongs. Value of Industry Shipments has been recoded to the following l-to-4 scale.

1 = $0 to $4.9 billion
2 = $5 billion to $13.9 billion
3 = $14 billion to $28.9 billion
4 = $29 billion or more

International Labor Database

This time-series database contains the civilian unemployment rates in percent from seven countries presented yearly over a 40-year period. The data are published by the Bureau of Labor Statistics of the U.S. Department of Labor. The countries are the United States, Canada, Australia, Japan, France, Germany, and Italy.

Financial Database

The financial database contains observations on eight variables for 100 companies. The variables are Type of Industry, Total Revenues ($ millions), Total Assets ($ millions), Return on Equity (%), Earnings per Share ($), Average Yield (%), Dividends per Share ($), and Average Price per Earnings (P/E) ratio. The companies represent seven different types of industries. The variable Type displays a company's industry type as:

1 = apparel
2 = chemical
3 = electric power
4 = grocery
5 = healthcare products
6 = insurance
7 = petroleum

Energy Database

The time-series energy database consists of data on five energy variables over a period of 26 years. The five variables are U.S. Energy Consumption, World Crude Oil Production, U.S. Nuclear Electricity Generation, U.S. Coal Production, and U.S. Natural Dry Gas Production. U.S. Energy Consumption is given in quadrillion BTUs per year, World Crude Oil Production is given in million barrels per day, U.S. Nuclear Electricity Generation is given in billion kilowatt-hours, U.S. Coal Production is given in million short tons, and U.S. Natural Dry Gas Production is given in million cubic feet.

U.S. and International Stock Market Database

This database contains seven variables—three from the U.S. stock market and four from international stock markets—with data representing monthly averages of each over a period of five years resulting in 60 data points per variable. The U.S. stock market variables include the Dow Jones Industrial Average, the NASDAQ, and Standard & Poor's 500. The four international stock market variables of Nikkei 225, Hang Seng, FTSE 100, and IPC represent Japan, Hong Kong, United Kingdom, and Mexico.

Hospital Database

This database contains observations for 11 variables on U.S. hospitals. These variables include Geographic Region, Control, Service, Number of Beds, Number of Admissions, Census, Number of Outpatients, Number of Births, Total Expenditures, Payroll Expenditures, and Personnel.

The region variable is coded from 1 to 7, and the numbers represent the following regions:

1 = South
2 = Northeast
3 = Midwest
4 = Southwest
5 = Rocky Mountain
6 = California
7 = Northwest

Control is a type of ownership. Four categories of control are included in the database:

1 = government, nonfederal
2 = nongovernment, not-for-profit
3 = for-profit
4 = federal government

Service is the type of hospital. The two types of hospitals used in this database are:

1 = general medical
2 = psychiatric

The total expenditures and payroll variables are in units of $1000.

Agribusiness Time-Series Database

The agribusiness time-series database contains the monthly weight (in 1000 lbs.) of cold storage holdings for six different vegetables and for total frozen vegetables over a 14-year period. Each of the seven variables represents 168 months of data. The six vegetables are green beans, broccoli, carrots, sweet corn, onions, and green peas. The data are published by the National Agricultural Statistics Service of the U.S. Department of Agriculture.

Assignment

Use the databases to answer the following questions.

1. In the manufacturing database, what is the level of data for each of the following variables?

- **a.** Number of Production Workers
- **b.** Cost of Materials
- **c.** Value of Industry Shipments
- **d.** Industry Group

2. In the hospital database, what is the level of data for each of the following variables?

- **a.** Region
- **b.** Control
- **c.** Number of Beds
- **d.** Personnel

3. In the financial database, what is the level of data for each of the following variables?

- **a.** Type of Industry
- **b.** Total Assets
- **c.** P/E Ratio

Case

Digiorno Pizza: Introducing a Frozen Pizza to Compete with Carry-Out

Kraft Foods successfully introduced DiGiorno Pizza into the marketplace in 1996, with first year sales of $120 million, followed by $200 million in sales in 1997. It was neither luck nor coincidence that DiGiorno Pizza was an instant success. Kraft conducted extensive research about the product and the marketplace before introducing this product to the public. Many questions had to be answered before Kraft began production. For example, why do people eat pizza? When do they eat pizza? Do consumers believe that carry-out pizza is always more tasty?

SMI-Alcott conducted a research study for Kraft in which they sent out 1000 surveys to pizza lovers. The results indicated that people ate pizza during fun social occasions or at home when no one wanted to cook. People used frozen pizza mostly for convenience but selected carry-out pizza for a variety of other reasons, including quality and the avoidance of cooking. The Loran Marketing Group conducted focus groups for Kraft with women aged 25 to 54. Their findings showed that consumers used frozen pizza for convenience but wanted carry-out pizza taste. Kraft researchers realized that if they were to launch a successful frozen pizza that could compete with carry-out pizza, they had to develop a frozen pizza that (a) had restaurant takeout quality, (b) possessed flavor variety, (c) was fast and easy to prepare, and (d) had the convenience of freezer storage. To satisfy these seemingly divergent goals, Kraft developed DiGiorno Pizza, which rises in the oven as it cooks. This impressed focus group members; and in a series of blind taste tests conducted by Product Dynamics, DiGiorno Pizza beat out all frozen pizzas and finished second overall behind one carry-out brand.

DiGiorno Pizza has continued to grow in sales and market share over the years. By 2005, sales had topped the $600 million mark, and DiGiorno Pizza held nearly a quarter of the market share of frozen pizza sales. In each of the last two quarters of 2009, DiGiorno sales increased 20%. On January 6, 2010, Kraft agreed to sell its North American frozen pizza business, including its DiGiorno products, to Nestlé for $3.7 billion. According to data reported by Statista, DiGiorno was by far the top frozen pizza brand in the United States in 2015 with over $957 million in sales when compared to the next brand which had sales of $463 million.

Discussion

Think about the market research that was conducted by Kraft and the fact that it used several companies.

1. What are some of the populations that Kraft might have been interested in measuring for these studies? Did Kraft actually attempt to contact entire populations? What samples were taken? In light of these two questions, how was the inferential process used by Kraft in their market research? Can you think of any descriptive statistics that might have been used by Kraft in their decision-making process?

2. In the various market research efforts made by Kraft for DiGiorno, some of the possible measurements appear in the following list. Categorize these by level of data. Think of some other measurements that Kraft researchers might have made to help them in this research effort, and categorize them by level of data.

   a. Number of pizzas consumed per week per household
   b. Age of pizza purchaser
   c. Zip code of the survey respondent
   d. Dollars spent per month on pizza per person
   e. Time in between purchases of pizza
   f. Rating of taste of a given pizza brand on a scale from 1 to 10, where 1 is very poor tasting and 10 is excellent taste
   g. Ranking of the taste of four pizza brands on a taste test
   h. Number representing the geographic location of the survey respondent
   i. Quality rating of a pizza brand as excellent, good, average, below average, poor
   j. Number representing the pizza brand being evaluated
   k. Sex of survey respondent

Source: Adapted from "Upper Crust," *American Demographics,* March 1999, p. 58; "Kraft Trading Pizza for Chocolate," *MarketWatch,* October 25, 1010, http://www.marketwatch.com/story/kraft-trading-pizza-for-chocolate-2010-01-05. "Sales of the leading 10 frozen pizza brands of the United States in 2015 (in million U.S. dollars)" at www.statista.com > Industries > Retail & Trade > Food & Beverage

Charts and Graphs

LEARNING OBJECTIVES

The overall objective of Chapter 2 is for you to master several techniques for summarizing and depicting data, thereby enabling you to:

1. Construct a frequency distribution from a set of data.
2. Construct different types of quantitative data graphs, including histograms, frequency polygons, ogives, dot plots, and stem-and-leaf plots, in order to interpret the data being graphed.
3. Construct different types of qualitative data graphs, including pie charts, bar graphs, and Pareto charts, in order to interpret the data being graphed.
4. Construct a cross-tabulation table and recognize basic trends in two-variable scatter plots of numerical data.

Decision Dilemma

Container Shipping Companies

Alberto Biscaro/Masterfile

For decades, businesspeople in many countries around the world wrestled with the issue of how to store and ship goods via trucks, trains, and ships. Various sizes and shapes of containers were developed to ship goods even within a country. The lack of consistent containers created a lot of extra work, as products were relocated from one container to another. Fortunately, in 1955 a former trucking company executive teamed up with an engineer to develop a version of the modern intermodal container that is widely used today. Because it is a standard size, this container in various forms can be moved from trucks to trains to ships without being opened, thereby eliminating the work of loading and unloading its contents multiple times. The International Organization for Standardization (ISO) has set up standards for the modern-day container, and perhaps the most commonly used container is 20 feet long and 8 feet wide. The container capacity of a ship is often measured in the number of 20-foot equivalent units or TEUs that can be loaded or unloaded from the vessel. Containerization has revolutionized cargo shipping, and today approximately 90% of non-bulk cargo worldwide moves by containers stacked on transport ships.

Shown in the next column are TEU capacities available on board-operated ships for the top five companies in the world as of October 21, 2015. Also included in the data is the total number of ships operated by each company.

COMPANY	TOTAL TEU CAPACITY	NUMBER OF SHIPS
APM-Maersk	3,031,701	594
Mediterranean Shipping Co.	2,660,981	497
CMA CGM Group	1,821,328	467
Evergreen Line	949,525	199
Hapag-Lloyd	924,417	174

MANAGERIAL AND STATISTICAL QUESTIONS

Suppose you are a shipping container industry analyst, and you are asked to prepare a brief report showing the leading shipping companies both in TEU shipping capacity and in number of ships.

1. What is the best way to display this shipping container company information? Are the raw data enough? Can you effectively display the data graphically?
2. Because some of the data are close together in size, is there a preferred graphical technique for differentiating between two or more similar numbers?

Source: "Alphaliner—TOP 100 Operated fleets as per 21 October 2015" at www.alphaliner.com/top100/

In Chapters 2 and 3, many techniques are presented for reformatting or reducing data so that the data are more manageable and can assist decision makers more effectively. Some of the most effective mechanisms for presenting data in a form meaningful to decision makers are graphical depictions. This chapter focuses on graphical tools for summarizing and presenting data. Through graphs and charts, the decision maker can often get an overall picture of the data and reach some useful conclusions merely by studying the chart or graph. Key characteristics of graphs often suggest appropriate choices among potential numerical methods (discussed in later chapters) for analyzing data. Visual representations of data are often much more effective communication tools than tables of numbers in business meetings.

A first step in exploring and analyzing data is to reduce important and sometimes expensive data to a graphic picture that is clear, concise and consistent with the message of the original data. Converting data to graphics can be creative and artful. In this chapter, guidelines are provided for selecting appropriate graphical representations for data sets. Charts and graphs discussed in Chapter 2 include histograms, frequency polygons, ogives, dot plots, stem-and-leaf plots, bar charts, pie charts, and Pareto charts for one-variable data and both cross-tabulation tables and scatter plots for two-variable numerical data.

2.1 Frequency Distributions

Raw data, or data that have not been summarized in any way, are sometimes referred to as **ungrouped data**. As an example, **Table 2.1** contains 60 years of raw data of the unemployment rates for Canada. *Data that have been organized into a frequency distribution* are called **grouped data**. **Table 2.2** presents a frequency distribution for the data displayed in Table 2.1.

TABLE 2.1 60 Years of Canadian Unemployment Rates (ungrouped data)

2.3	7.0	6.3	11.3	9.6
2.8	7.1	5.6	10.6	9.1
3.6	5.9	5.4	9.7	8.3
2.4	5.5	7.1	8.8	7.6
2.9	4.7	7.1	7.8	6.8
3.0	3.9	8.0	7.5	7.2
4.6	3.6	8.4	8.1	7.7
4.4	4.1	7.5	10.3	7.6
3.4	4.8	7.5	11.2	7.2
4.6	4.7	7.6	11.4	6.8
6.9	5.9	11.0	10.4	6.3
6.0	6.4	12.0	9.5	6.0

TABLE 2.2 Frequency Distribution of 60 Years of Unemployment Data for Canada (grouped data)

CLASS INTERVAL	FREQUENCY
1–under 3	4
3–under 5	12
5–under 7	13
7–under 9	19
9–under 11	7
11–under 13	5

The distinction between ungrouped and grouped data is important because the calculation of statistics differs between the two types of data. Several of the charts and graphs presented in this chapter are constructed from grouped data.

One particularly useful tool for grouping data is the **frequency distribution**, which is *a summary of data presented in the form of class intervals and frequencies.* How is a frequency distribution constructed from raw data? That is, how are frequency distributions like the one displayed in Table 2.2 constructed from raw data like those presented in Table 2.1? Frequency distributions are relatively easy to construct. Although some guidelines and rules of thumb help in their construction, frequency distributions vary in final shape and design, even when the original raw data are identical. In a sense, frequency distributions are constructed according to individual business researchers' taste.

When constructing a frequency distribution, the business researcher should first determine the range of the raw data. The **range** often is defined as *the difference between the largest and smallest numbers.* The range for the data in Table 2.1 is 9.7 (12.0–2.3).

The second step in constructing a frequency distribution is to determine how many classes it will contain. One rule of thumb is to select between *5 and 15 classes*. If the frequency distribution contains too few classes, the data summary may be too general to be useful. Too many classes may result in a frequency distribution that does not aggregate the data enough to be helpful. The final number of classes is arbitrary. The business researcher arrives at a number by examining the range and determining a number of classes that will span the range adequately and also be meaningful to the user. The data in Table 2.1 were grouped into six classes for Table 2.2.

After selecting the number of classes, the business researcher must determine the width of the class interval. An approximation of the class width can be calculated by dividing the range by the number of classes. For the data in Table 2.1, this approximation would be $9.7/6 = 1.62$. Normally, the number is rounded up to the next whole number, which in this case is 2. The frequency distribution must start at a value equal to or lower than the lowest number of the ungrouped data and end at a value equal to or higher than the highest number. The lowest unemployment rate is 2.3 and the highest is 12.0, so the business researcher starts the frequency distribution at 1 and ends it at 13. Table 2.2 contains the completed frequency distribution for the data in Table 2.1. Class endpoints are selected so that no value of the data can fit into more than one class. The class interval expression "under" in the distribution of Table 2.2 avoids such a problem.

Class Midpoint

The *midpoint of each class interval* is called the **class midpoint** and is sometimes referred to as the **class mark**. It is *the value halfway across the class interval* and can be calculated as *the average of the two class endpoints*. For example, in the distribution of Table 2.2, the midpoint of the class interval 3–under 5 is 4, or $(3 + 5)/2$.

The class midpoint is important, because it becomes the representative value for each class in most group statistics calculations. (Chapter 3) The third column in **Table 2.3** contains the class midpoints for all classes of the data from Table 2.2.

TABLE 2.3 Class Midpoints, Relative Frequencies, and Cumulative Frequencies for Unemployment Data

INTERVAL	FREQUENCY	CLASS MIDPOINT	RELATIVE FREQUENCY	CUMULATIVE FREQUENCY
1–under 3	4	2	.0667	4
3–under 5	12	4	.2000	16
5–under 7	13	6	.2167	29
7–under 9	19	8	.3167	48
9–under 11	7	10	.1167	55
11–under 13	5	12	.0833	60
Total	60			

Relative Frequency

Relative frequency is *the proportion of the total frequency that is in any given class interval in a frequency distribution*. Relative frequency is the individual class frequency divided by the total frequency. For example, from Table 2.3, the relative frequency for the class interval 5–under 7 is $13/60 = .2167$. Consideration of the relative frequency is preparatory to the study of probability in Chapter 4. Indeed, if values were selected randomly from the data in Table 2.1, the probability of drawing a number that is "5–under 7" would be .2167, the relative frequency for that class interval. The fourth column of Table 2.3 lists the relative frequencies for the frequency distribution of Table 2.2.

Cumulative Frequency

The **cumulative frequency** is *a running total of frequencies through the classes of a frequency distribution*. The cumulative frequency for each class interval is the frequency for that class interval added to the preceding cumulative total. In Table 2.3, the cumulative frequency for the first class is the same as the class frequency: 4. The cumulative frequency for the second class interval is the frequency of that interval (12) plus the frequency of the first interval (4), which yields a new cumulative frequency of 16. This process continues through the last interval, at which point the cumulative total equals the sum of the frequencies (60). The concept of cumulative frequency is used in many areas, including sales cumulated over a fiscal year, sports scores during a contest (cumulated points), years of service, points earned in a course, and costs of doing business over a period of time. Table 2.3 gives cumulative frequencies for the data in Table 2.2.

DEMONSTRATION PROBLEM 2.1

The following data from the Federal Home Loan Mortgage Corporation are the average monthly 30-year fixed rate mortgage interest rates for a recent 40-month period.

5.06	4.89	4.75	4.11	3.81
4.95	4.74	4.95	4.07	3.69
4.88	4.56	4.83	3.99	3.55
4.88	4.43	4.84	3.95	3.60
5.05	4.35	4.64	3.92	3.50
4.99	4.24	4.51	3.89	3.38
4.97	4.30	4.54	3.95	3.35
5.10	4.71	4.27	3.91	3.34

Construct a frequency distribution for these data. Calculate and display the class midpoints, relative frequencies, and cumulative frequencies for this frequency distribution.

Solution: How many classes should this frequency distribution contain? The range of the data is 1.76 (5.10 – 3.34). If 8 classes are used, each class width is approximately:

$$\text{Class Width} = \frac{\text{Range}}{\text{Number of Classes}} = \frac{1.76}{8} = 0.22$$

If a class width of .25 is used, a frequency distribution can be constructed with endpoints that are more uniform looking and allow for presentation of the information in categories more familiar to mortgage interest rate users.

The first endpoint must be 3.34 or lower to include the smallest value; the last endpoint must be 5.10 or higher to include the largest value. In this case, the frequency distribution begins at 3.25 and ends at 5.25. The resulting frequency distribution, class midpoints, relative frequencies, and cumulative frequencies are listed in the following table:

INTERVAL	FREQUENCY	CLASS MIDPOINT	RELATIVE FREQUENCY	CUMULATIVE FREQUENCY
3.25–under 3.50	3	3.375	.075	3
3.50–under 3.75	4	3.625	.100	7
3.75–under 4.00	7	3.875	.175	14
4.00–under 4.25	3	4.125	.075	17
4.25–under 4.50	4	4.375	.100	21
4.50–under 4.75	6	4.625	.150	27
4.75–under 5.00	10	4.875	.250	37
5.00–under 5.25	3	5.125	.075	40
Total	40			

The frequencies and relative frequencies of these data reveal the mortgage interest rate classes that are likely to occur during this period of time. Overall, the mortgage rates are distributed relatively evenly, with the 4.75–under 5.00 class interval containing the greatest frequency (10), followed by the 3.75–under 4.00 class interval (7), and the 4.50–under 4.75 interval (6).

2.1 Problems

2.1. The following data represent the afternoon high temperatures for 50 construction days during a year in St. Louis.

42	70	64	47	66	69	73	38	48	25
55	85	10	24	45	31	62	47	63	84
16	40	81	15	35	17	40	36	44	17
38	79	35	36	23	64	75	53	31	60
31	38	52	16	81	12	61	43	30	33

a. Construct a frequency distribution for the data using five class intervals.

b. Construct a frequency distribution for the data using 10 class intervals.

c. Examine the results of (a) and (b) and comment on the usefulness of the frequency distribution in terms of temperature summarization capability.

2.2. A packaging process is supposed to fill small boxes of raisins with approximately 50 raisins so that each box will weigh the same. However, the number of raisins in each box will vary. Suppose 100 boxes of raisins are randomly sampled, the raisins counted, and the following data are obtained.

57	51	53	52	50	60	51	51	52	52
44	53	45	57	39	53	58	47	51	48
49	49	44	54	46	52	55	54	47	53
49	52	49	54	57	52	52	53	49	47
51	48	55	53	55	47	53	43	48	46
54	46	51	48	53	56	48	47	49	57
55	53	50	47	57	49	43	58	52	44
46	59	57	47	61	60	49	53	41	48
59	53	45	45	56	40	46	49	50	57
47	52	48	50	45	56	47	47	48	46

a. Construct a frequency distribution for these data.

b. What does the frequency distribution reveal about the box fills?

2.3. The owner of a fast-food restaurant ascertains the ages of a sample of customers. From these data, the owner constructs the frequency distribution shown. For each class interval of the frequency distribution, determine the class midpoint, the relative frequency, and the cumulative frequency.

CLASS INTERVAL	FREQUENCY
0–under 5	6
5–under 10	8
10–under 15	17
15–under 20	23
20–under 25	18
25–under 30	10
30–under 35	4

What does the relative frequency tell the fast-food restaurant owner about customer ages?

2.4. The human resources manager for a large company commissions a study in which the employment records of 500 company employees are examined for absenteeism during the past year. The business researcher conducting the study organizes the data into a frequency distribution to assist the human resources manager in analyzing the data. The frequency distribution is shown. For each class of the frequency distribution, determine the class midpoint, the relative frequency, and the cumulative frequency.

CLASS INTERVAL	FREQUENCY
0–under 2	218
2–under 4	207
4–under 6	56
6–under 8	11
8–under 10	8

2.5. List three specific uses of cumulative frequencies in business.

2.2 Quantitative Data Graphs

Interactive Applet

One of the most effective mechanisms for presenting data in a form meaningful to decision makers is graphical depiction. Through graphs and charts, the decision maker can often get an overall picture of the data and reach some useful conclusions merely by studying the chart or graph. Converting data to graphics can be creative and artful. Often the most difficult step in this process is to reduce data to a graphic picture that is both clear and concise and yet consistent with the message of the original data. One of the most important uses of graphical depiction in statistics is to help the researcher determine the shape of a distribution. Data graphs can generally be classified as quantitative or qualitative. Quantitative data graphs are plotted along a numerical scale, and qualitative graphs are plotted using non-numerical categories. In this section, we will examine five types of quantitative data graphs: (1) histogram, (2) frequency polygon, (3) ogive, (4) dot plot, and (5) stem-and-leaf plot.

Histograms

One of the more widely used types of graphs for quantitative data is the **histogram**. A histogram is a series of contiguous rectangles that represent the frequency of data in given class intervals. If the class intervals used along the horizontal axis are equal, then the heights of the rectangles represent the frequency of values in a given class interval. If the class intervals are unequal, then the areas of the rectangles can be used for relative comparisons of class frequencies. Construction of a histogram involves labeling the x-axis (abscissa) with the class endpoints and the y-axis (ordinate) with the frequencies, drawing a horizontal line segment from class endpoint to class endpoint at each frequency value, and connecting each line segment vertically from the frequency value to the x-axis to form a series of rectangles. **Figure 2.1** is a histogram of the frequency distribution in Table 2.2 produced by using the software package Minitab.

A histogram is a useful tool for differentiating the frequencies of class intervals. A quick glance at a histogram reveals which class intervals produce the highest frequency totals. Figure 2.1 clearly shows that the class interval 7–under 9 yields by far the highest frequency count (19). Examination of the histogram reveals where large increases or decreases occur between classes, such as from the 1–under 3 class to the 3–under 5 class, an increase of 8, and from the 7–under 9 class to the 9–under 11 class, a decrease of 12.

Note that the scales used along the x- and y-axes for the histogram in Figure 2.1 are almost identical. However, because ranges of meaningful numbers for the two variables being graphed often differ considerably, the graph may have different scales on the two

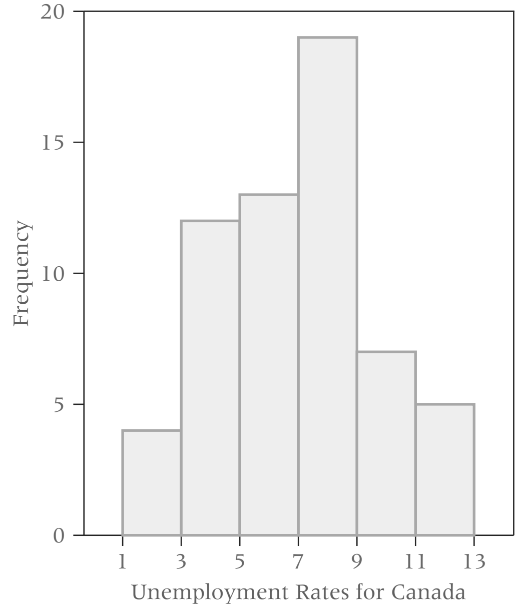


FIGURE 2.1 Minitab Histogram of Canadian Unemployment Data

axes. Figure 2.2 shows what the histogram of unemployment rates would look like if the scale on the *y*-axis were more compressed than that on the *x*-axis. Notice that with the compressed graph, Figure 2.2, there appears to be less difference between the lengths of the rectangles than those in Figure 2.1 implying that the differences in frequencies for the compressed graph are not as great as they are in Figure 2.1. It is important that the user of the graph clearly understands the scales used for the axes of a histogram. Otherwise, a graph's creator can "lie with statistics" by stretching or compressing a graph to make a point.*

Using Histograms to Get an Initial Overview of the Data Because of the widespread availability of computers and statistical software packages to business researchers and decision makers, the histogram continues to grow in importance in yielding information about the shape of the distribution of a large database, the variability of the data, the central location of the data, and outlier data. Although most of these concepts are presented in Chapter 3, the notion of histogram as an initial tool to access these data characteristics is presented here.

A business researcher measured the volume of stocks traded on Wall Street three times a month for nine years, resulting in a database of 324 observations. Suppose a financial decision maker wants to use these data to reach some conclusions about the stock market. Figure 2.3

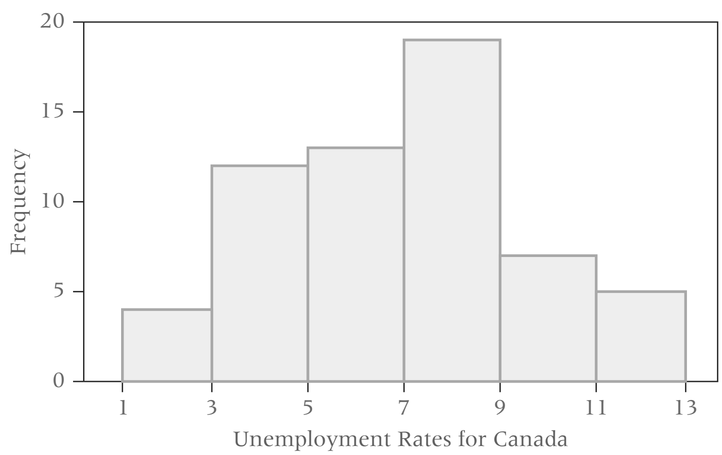


FIGURE 2.2 Minitab Histogram of Canadian Unemployment Data (*y*-axis compressed)

*It should be pointed out that the software package Excel uses the term *histogram* to refer to a frequency distribution. However, by checking Chart Output in the Excel histogram dialog box, a graphical histogram is also created.

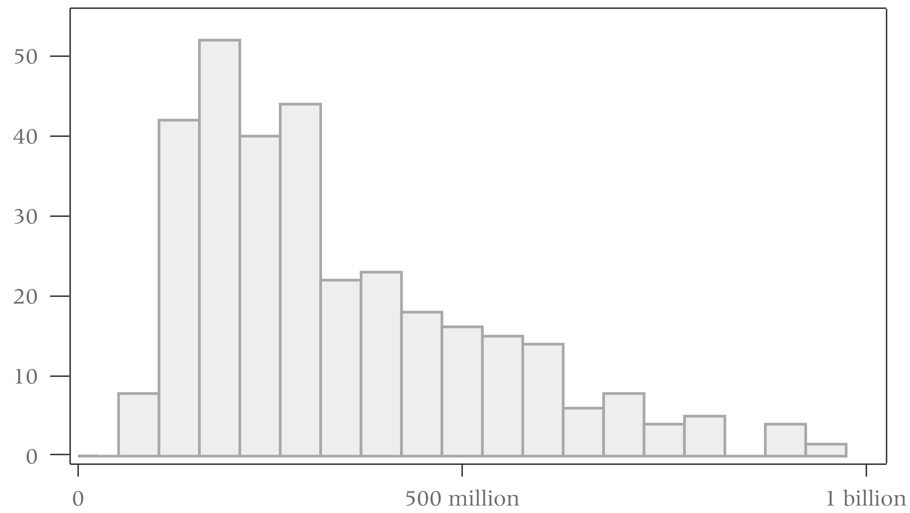


FIGURE 2.3 **Histogram of Stock Volumes**

shows a Minitab-produced histogram of these data. What can we learn from this histogram? Virtually all stock market volumes fall between zero and 1 billion shares. The distribution takes on a shape that is high on the left end and tapered to the right. In Chapter 3 we will learn that the shape of this distribution is skewed toward the right end. In statistics, it is often useful to determine whether data are approximately normally distributed (bell-shaped curve) as shown in **Figure 2.4**. We can see by examining the histogram in Figure 2.3 that the stock market volume data are not normally distributed. Although the center of the histogram is located near 500 million shares, a large portion of stock volume observations falls in the lower end of the data somewhere between 100 million and 400 million shares. In addition, the histogram shows some outliers in the upper end of the distribution. Outliers are data points that appear outside of the main body of observations and may represent phenomena that differ from those represented by other data points. By observing the histogram, we notice a few data observations near 1 billion. One could conclude that on a few stock market days an unusually large volume of shares is traded. These and other insights can be gleaned by examining the histogram and show that histograms play an important role in the initial analysis of data.

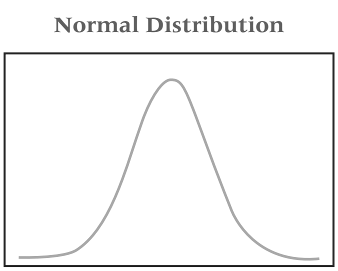


FIGURE 2.4 **Normal Distribution**

Frequency Polygons

A **frequency polygon**, like the histogram, is a graphical display of class frequencies. However, instead of using rectangles like a histogram, in a frequency polygon each class frequency is plotted as a dot at the class midpoint, and the dots are connected by a series of line segments. Construction of a frequency polygon begins by scaling class midpoints along the horizontal axis and the frequency scale along the vertical axis. A dot is plotted for the associated frequency value at each class midpoint. Connecting these midpoint dots completes the graph. **Figure 2.5** shows a frequency polygon of the distribution data from Table 2.2 produced by using the software package Excel. The information gleaned from frequency polygons and histograms is similar. As with the histogram, changing the scales of the axes can compress or stretch a frequency polygon, which affects the user's impression of what the graph represents.

Ogives

An **ogive** (o-jive) is *a cumulative frequency polygon*. Construction begins by labeling the x-axis with the class endpoints and the y-axis with the frequencies. However, the use of cumulative frequency values requires that the scale along the y-axis be great enough to include

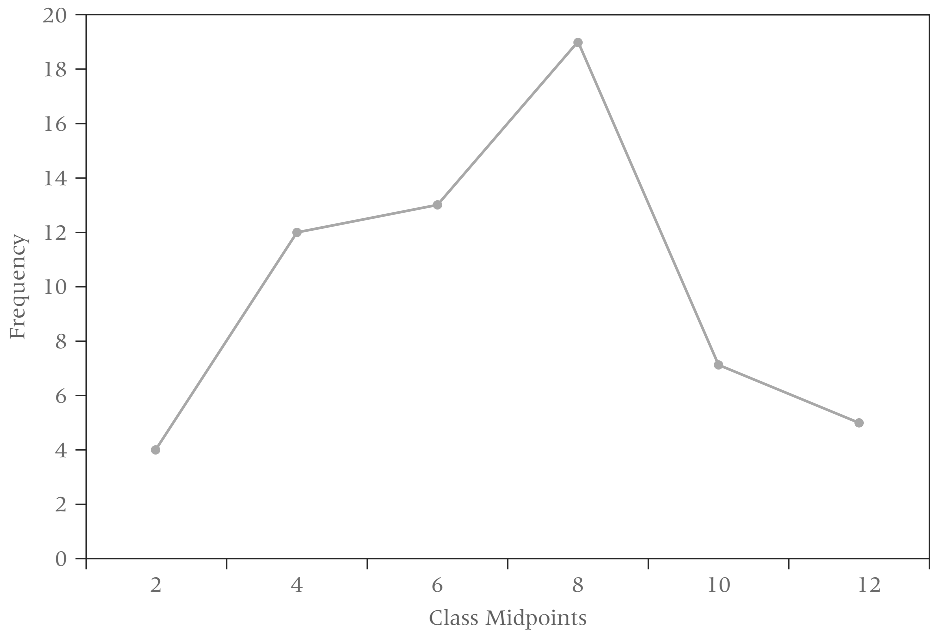


FIGURE 2.5 **Excel-Produced Frequency Polygon of the Unemployment Data**

the frequency total. A dot of zero frequency is plotted at the beginning of the first class, and construction proceeds by marking a dot at the *end* of each class interval for the cumulative value. Connecting the dots then completes the ogive. Figure 2.6 presents an ogive produced by using Excel for the data in Table 2.2.

Ogives are most useful when the decision maker wants to see *running totals.* For example, if a comptroller is interested in controlling costs, an ogive could depict cumulative costs over a fiscal year.

Steep slopes in an ogive can be used to identify sharp increases in frequencies. In Figure 2.6, a particularly steep slope occurs in the 7–under 9 class, signifying a large jump in class frequency totals.

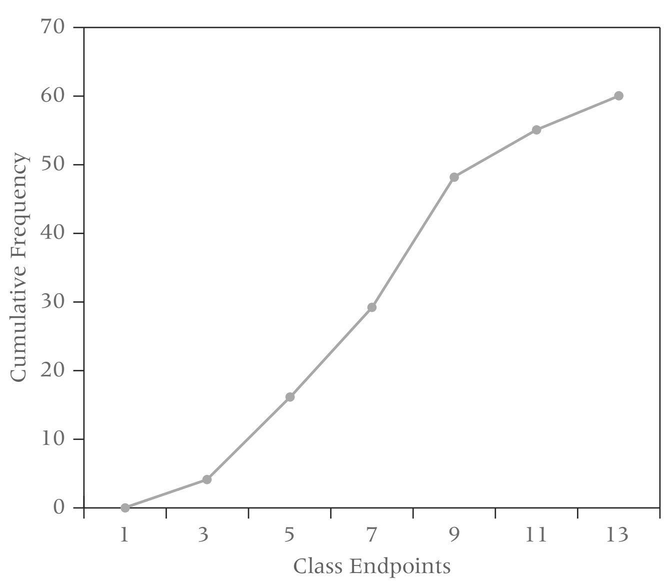


FIGURE 2.6 **Excel Ogive of the Unemployment Data**

Dot Plots

A relatively simple statistical chart that is generally used to display continuous, quantitative data is the **dot plot**. In a dot plot, each data value is plotted along the horizontal axis and is represented on the chart by a dot. If multiple data points have the same values, the dots will stack up vertically. If there are a large number of close points, it may not be possible to display all of the data values along the horizontal axis. Dot plots can be especially useful for observing the overall shape of the distribution of data points along with identifying data values or intervals for which there are groupings and gaps in the data. **Figure 2.7** displays a minitab-produced dot plot for the Canadian unemployment data shown in Table 2.1. Note that the distribution is relatively balanced with a peak near the center. There are a few gaps to note, such as from 4.9 to 5.3, from 9.9 to 10.2, and from 11.5 to 11.9. In addition, there are groupings around 6.0, 7.1, and 7.5.

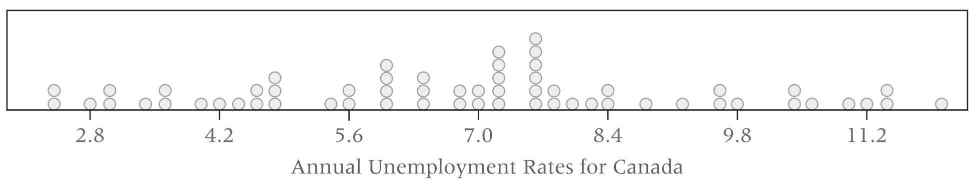


FIGURE 2.7 **A Minitab-Produced Dot Plot of the Canadian Unemployment Data**

Stem-and-Leaf Plots

Applied Skills Video

Another way to organize raw data into groups besides using a frequency distribution is a **stem-and-leaf plot**. This technique is simple and provides a unique view of the data. A stem-and-leaf plot is constructed by separating the digits for each number of the data into two groups, *a stem and a leaf*. The leftmost digits are the stem and consist of the higher valued digits. The rightmost digits are the leaves and contain the lower values. If a set of data has only two digits, the stem is the value on the left and the leaf is the value on the right. For example, if 34 is one of the numbers, the stem is 3 and the leaf is 4. For numbers with more than two digits, division of stem and leaf is a matter of researcher preference.

Table 2.4 contains scores from an examination on plant safety policy and rules given to a group of 35 job trainees. A stem-and-leaf plot of these data is displayed in **Table 2.5**. One advantage of such a distribution is that the instructor can readily see whether the scores are in the upper or lower end of each bracket and also determine the spread of the scores. A second advantage of stem-and-leaf plots is that the values of the original raw data are retained (whereas most frequency distributions and graphic depictions use the class midpoint to represent the values in a class).

TABLE 2.4 **Safety Examination Scores for Plant Trainees**

86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	74	39	67
79	83	70	78	91
68	49	56	94	81

TABLE 2.5 Stem-and-Leaf Plot for Plant Safety Examination Data

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

DEMONSTRATION PROBLEM 2.2

The following data represent the costs (in dollars) of a sample of 30 postal mailings by a company.

3.67	2.75	9.15	5.11	3.32	2.09
1.83	10.94	1.93	3.89	7.20	2.78
6.72	7.80	5.47	4.15	3.55	3.53
3.34	4.95	5.42	8.64	4.84	4.10
5.10	6.45	4.65	1.97	2.84	3.21

Using dollars as a stem and cents as a leaf, construct a stem-and-leaf plot of the data.

Solution:

STEM	LEAF						
1	83	93	97				
2	09	75	78	84			
3	21	32	34	53	55	67	89
4	10	15	65	84	95		
5	10	11	42	47			
6	45	72					
7	20	80					
8	64						
9	15						
10	94						

2.2 Problems

2.6. Assembly times for components must be understood in order to "level" the stages of a production process. Construct both a histogram and a frequency polygon for the following assembly time data and comment on the key characteristics of the distribution.

CLASS INTERVAL	FREQUENCY
30–under 32	5
32–under 34	7
34–under 36	15
36–under 38	21
38–under 40	34
40–under 42	24
42–under 44	17
44–under 46	8

2.7. A call center is trying to better understand staffing requirements. It investigates the number of calls received during the evening shift and obtains the information given below. Construct a histogram of the data and comment on the key characteristics of the distribution. Construct a frequency polygon and compare it to the histogram. Which do you prefer, and why?

CLASS INTERVAL	FREQUENCY
10–under 20	9
20–under 30	7
30–under 40	10
40–under 50	6
50–under 60	13
60–under 70	18
70–under 80	15

2.8. Construct an ogive for the following data.

CLASS INTERVAL	FREQUENCY
3–under 6	2
6–under 9	5
9–under 12	10
12–under 15	11
15–under 18	17
18–under 21	5

2.9. A real estate group is investigating the price a condominiums of a given size (sq ft). The following sales prices ($1,000) were obtained in one region of a city. Construct a stem-and-leaf plot for the following data using two digits for the stem. Comment on the key characteristics of the distribution. Construct a dot plot of the data and comment on how it differs from the stem-and-leaf plot in providing information about the data.

212	239	240	218	222	249	265	224
257	271	266	234	239	219	255	260
243	261	249	230	246	263	235	229
218	238	254	249	250	263	229	221
253	227	270	257	261	238	240	239
273	220	226	239	258	259	230	262
255	226						

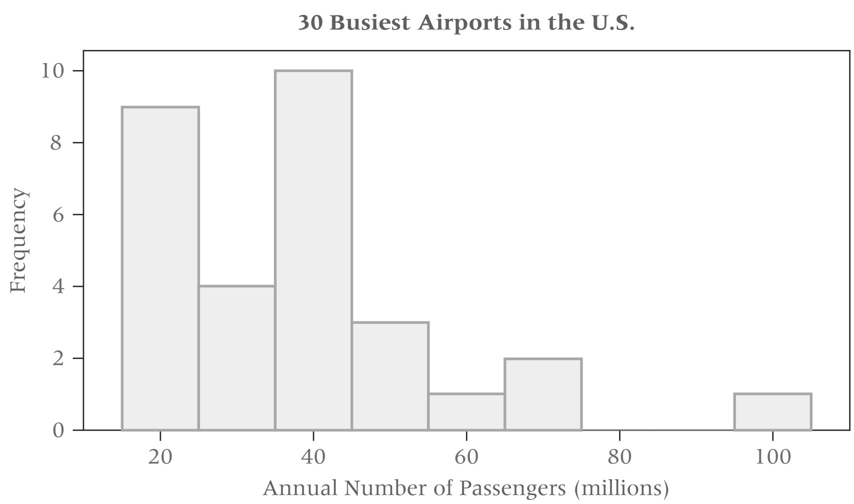
2.10. The following data represent the number of passengers per flight in a sample of 50 flights from Wichita, Kansas, to Kansas City, Missouri.

23	46	66	67	13	58	19	17	65	17
25	20	47	28	16	38	44	29	48	29
69	34	35	60	37	52	80	59	51	33
48	46	23	38	52	50	17	57	41	77
45	47	49	19	32	64	27	61	70	19

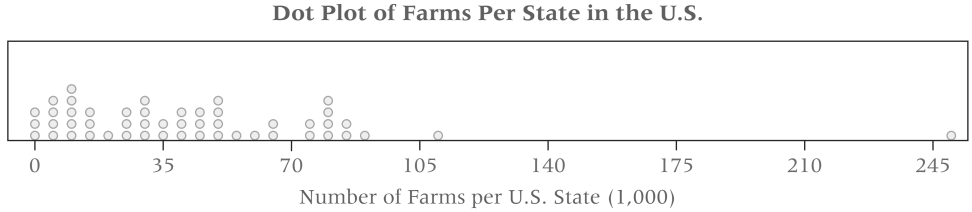
a. Construct a dot plot for these data.

b. Construct a stem-and-leaf plot for these data. What does the stem-and-leaf plot tell you about the number of passengers per flight?

2.11. The Airports Council International—North America (ACI) publishes data on the busiest airports in the United States. Shown below is a Minitab-produced histogram constructed from ACI data on the number of passengers that enplaned and deplaned in 2014 in the United States. As an example, Atlanta's Hartsfield–Jackson International Airport was the busiest airport in the United States with 96,178,899 passengers. What are some observations that you can make from the graph? Describe the top 30 busiest airports in the United States using this histogram.



2.12. Study the Minitab-produced dot plot of the number of farms per state in the United States shown below. Comment on any observations that you make from the graph. What does this graph tell you about the number of farms per state? The average number of farms per state calculated from the raw data (not given here) and sourced from the U.S. Department of Agriculture is 44,060. Reconcile this number with the dot plot.



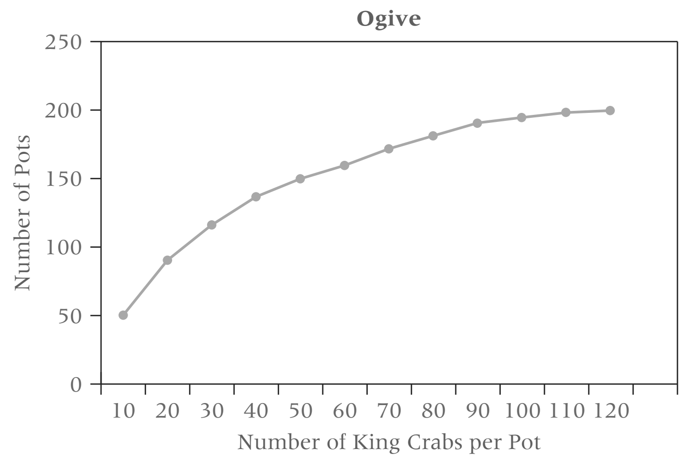
2.13. A full-service car wash has an automated exterior conveyor car wash system that does the initial cleaning in a few minutes. However, once the car is through the system, car wash workers hand clean the inside and the outside of the car for approximately 15 to 25 additional minutes. There are enough workers to handle four cars at once during this stage. On a busy day with good weather, the car wash can handle up to 150 cars in a 12-hour time period. However, on rainy days or on certain days of the year, business is slow. Suppose 50 days of work are randomly sampled from the car wash's records and the number of cars washed each day is recorded. A stem-and-leaf plot of this output is constructed and is given below. Study the plot and write a few sentences describing the number of cars washed per day over this period of work. Note that the stem-and-leaf display is from Minitab, the stems are in the middle column, each leaf is only one digit and is shown in the right column, and the numbers in the left column are cumulative frequencies up to the median and then decumulative thereafter.

Stem-and-Leaf Display: Cars Washed Per Day

Stem-and-leaf of Cars Washed Per Day N = 50 Leaf Unit = 1.0

	STEM	LEAF
3	2	599
9	3	344778
15	4	015689
18	5	378
21	6	223
24	7	457
(3)	8	112
23	9	05
21	10	1234578
14	11	466
11	12	01467
6	13	37
4	14	1457

2.14. A hundred or so boats go fishing every year for three or four weeks off of the Bering Strait for Alaskan king crabs. To catch these king crabs, large pots are baited and left on the sea bottom, often several hundred feet deep. Because of the investment in boats, equipment, personnel, and supplies, fishing for such crabs can be financially risky if not enough crabs are caught. Thus, as pots are pulled and emptied, there is great interest in how many legal king crabs (males of a certain size) there are in any given pot. Suppose the number of legal king crabs is reported for each pot during a season and recorded. In addition, suppose that 200 of these are randomly selected and the numbers per pot are used to create the ogive shown below. Study the ogive and comment on the number of legal king crabs per pot.



2.3 Qualitative Data Graphs

In contrast to quantitative data graphs that are plotted along a numerical scale, qualitative graphs are plotted using non-numerical categories. In this section, we will examine three types of qualitative data graphs: (1) pie charts, (2) bar charts, and (3) Pareto charts.

Pie Charts

A **pie chart** is *a circular depiction of data where the area of the whole pie represents 100% of the data and slices of the pie represent a percentage breakdown of the sublevels*. Pie charts show the relative magnitudes of the parts to the whole. They are widely used in business, particularly to depict such things as budget categories, market share, and time/resource allocations. However, the use of pie charts is minimized in the sciences and technology because pie charts can lead to less accurate judgments than are possible with other types of graphs.* Generally, it is more difficult for the viewer to interpret the relative size of angles in a pie chart than to judge the length of rectangles in a bar chart. In the feature Thinking Critically about Statistics in Business Today, "Where Are Soft Drinks Sold?" the percentage of sales by place are displayed by both a pie chart and a vertical bar chart.

Construction of the pie chart begins by determining the proportion of the subunit to the whole. **Table 2.6** contains the refining capacity (1,000 barrels per day) of the

*William S. Cleveland, *The Elements of Graphing Data*. Monterey, CA: Wadsworth Advanced Books and Software, 1985.

TABLE 2.6 Top Five U.S. Petroleum Refining Companies in Capacity

COMPANY	CAPACITY (1,000 BARRELS PER DAY)	PROPORTION	DEGREES
Exxon Mobil	5,589	.3693	132.95
Valero Energy	2,777	.1835	66.06
Chevron	2,540	.1678	60.41
ConocoPhillips	2,514	.1661	59.79
Marathon Oil	1,714	.1133	40.79
Totals	15,134	1.0000	360.00

top five petroleum refining companies in the United States in a recent year. To construct a pie chart from these data, first convert the raw capacity figures to proportions by dividing each capacity figure by the total capacity figure (15,134). This proportion is analogous to the relative frequency computed for frequency distributions. Because a circle contains 360°, each proportion is then multiplied by 360 to obtain the number of degrees to represent each company in the pie chart. For example, Exxon Mobil's capacity of 5,589 (1,000 barrels) represents a .3693 proportion of the total capacity for these five companies. $\left(\frac{5{,}589}{15{,}134} = .3693\right)$. Multiplying this value by 360° results in an angle of 132.95°. The pie chart is then constructed by determining each of the other angles and using a compass to lay out the slices. The pie chart in Figure 2.8, constructed by using Minitab, depicts the data from Table 2.6.

Bar Graphs

Another widely used qualitative data graphing technique is the **bar graph** or **bar chart**. A bar graph or chart contains two or more categories along one axis and a series of bars, one for each category, along the other axis. Typically, the length of the bar represents the magnitude of the measure (amount, frequency, money, percentage, etc.) for each category. The bar graph is qualitative because the categories are non-numerical, and it may be either horizontal or vertical. In Excel, horizontal bar graphs are referred to as **bar charts**, and vertical bar graphs are referred to as **column charts**. A bar graph is generally constructed from the same type of data that is used to produce a pie chart. However, an advantage of using a bar graph over a pie

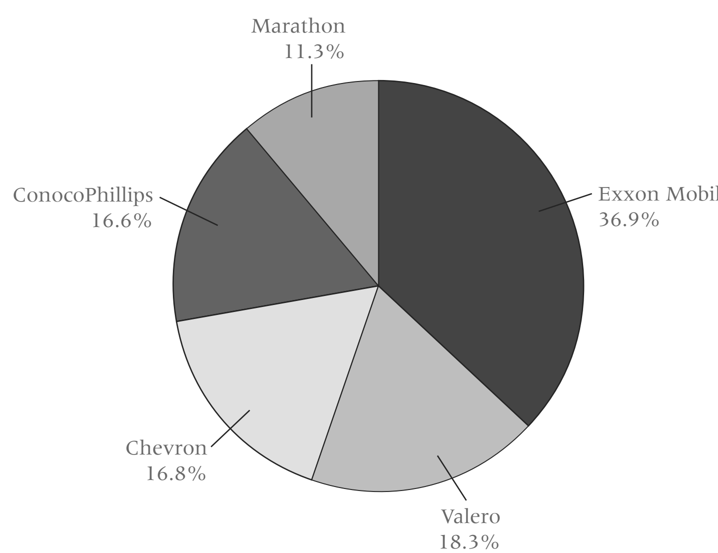


FIGURE 2.8 Minitab Pie Chart of the Petroleum Refining Capacity of the Top Five U.S. Companies

TABLE 2.7 How Much is Spent on Back-to-College Shopping by the Average Student

CATEGORY	AMOUNT SPENT ($ US)
Electronics	$211.89
Clothing and Accessories	134.40
Dorm Furnishings	90.90
School Supplies	68.47
Misc.	93.72

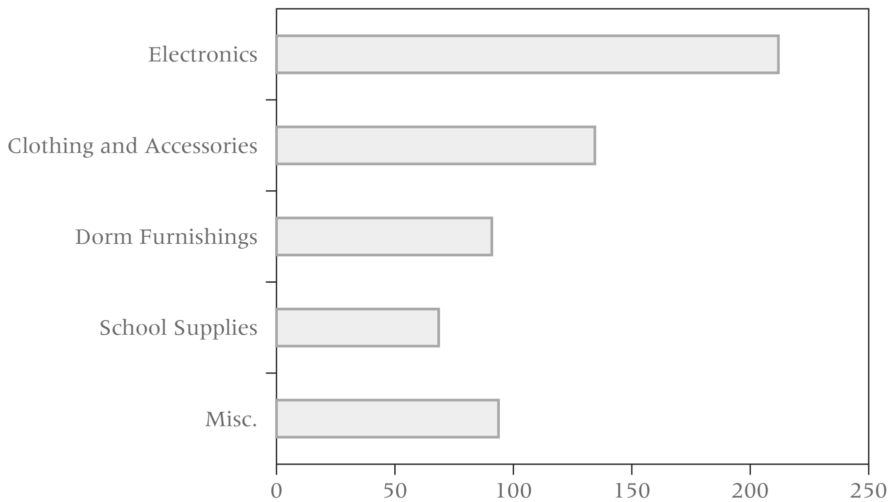


FIGURE 2.9 Bar Graph of Back-to-College Spending

chart for a given set of data is that for categories that are close in value, it is considered easier to see the difference in the bars of bar graph than discriminating between pie slices.

As an example, consider the data in Table 2.7 regarding how much the average college student spends on back-to-college spending. Constructing a bar graph from these data, the categories are Electronics, Clothing and Accessories, Dorm Furnishings, School Supplies, and Misc. Bars for each of these categories are made using the dollar figures given in the table. The resulting bar graph produced by Excel is shown in Figure 2.9.

DEMONSTRATION PROBLEM 2.3

According to the National Retail Federation and Center for Retailing Education at the University of Florida, the four main sources of inventory shrinkage are employee theft, shoplifting, administrative error, and vendor fraud. The estimated annual dollar amount in shrinkage ($ millions) associated with each of these sources follows:

Employee theft	$17,918.6
Shoplifting	15,191.9
Administrative error	7,617.6
Vendor fraud	2,553.6
Total	$43,281.7

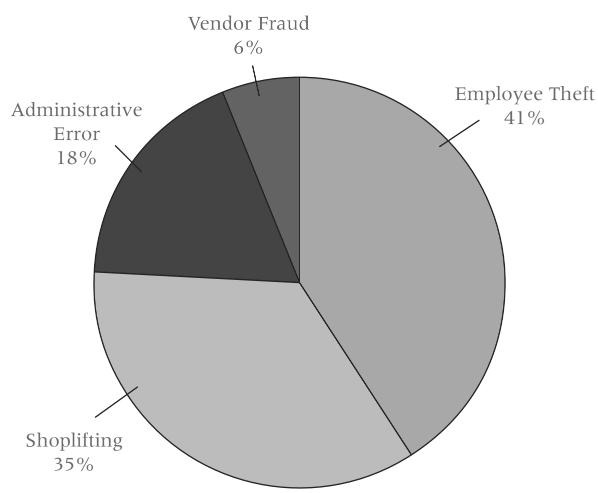
Construct a pie chart and a bar chart to depict these data.

Solution: To produce a pie chart, convert each raw dollar amount to a proportion by dividing each individual amount by the total.

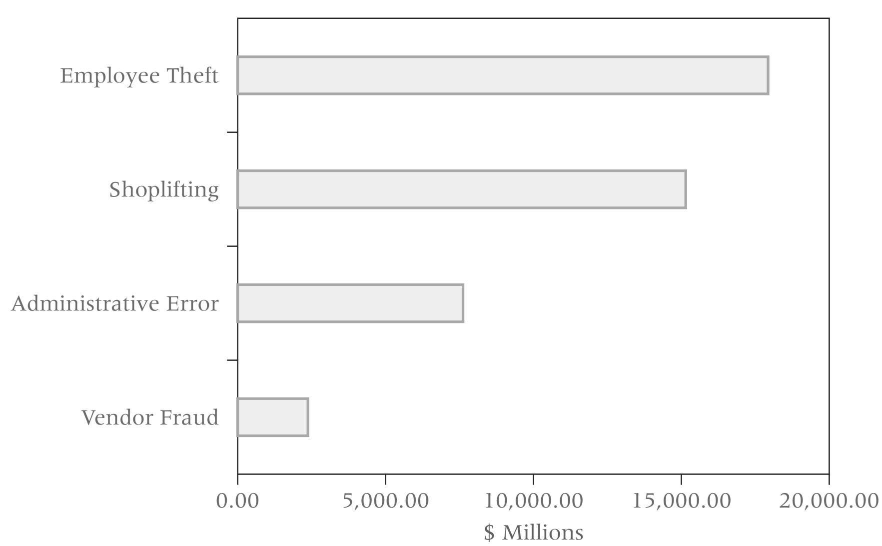
Employee theft	17,918.6/43,281.7 = .414
Shoplifting	15,191.9/43,281.7 = .351
Administrative error	7,617.6/43,281.7 = .176
Vendor fraud	2,553.6/43,281.7 = .059
Total	1.000

Convert each proportion to degrees by multiplying each proportion by 360°.

Employee theft	.414 · 360° = 149.0°
Shoplifting	.351 · 360° = 126.4°
Administrative error	.176 · 360° = 63.4°
Vendor fraud	.059 · 360° = 21.2°
Total	360.0°



Using the raw data above, we can produce the following bar chart.



Pareto Charts

A third type of qualitative data graph is a Pareto chart, which can be viewed as a particular application of the bar graph. An important concept and movement in business is total quality management (see Chapter 18). One of the key aspects of total quality management is the constant search for causes of problems in products and processes. A graphical technique for displaying problem causes is Pareto analysis. Pareto analysis is a quantitative tallying of the number and types of defects that occur with a product or service. Analysts use this tally to produce *a vertical bar chart that displays the most common types of defects, ranked in order of occurrence from left to right*. The bar chart is called a **Pareto chart**.

Pareto charts were named after an Italian economist, Vilfredo Pareto, who observed more than 100 years ago that most of Italy's wealth was controlled by a few families who were the major drivers behind the Italian economy. Quality expert J. M. Juran applied this notion to the quality field by observing that poor quality can often be addressed by attacking a few major causes that result in most of the problems. A Pareto chart enables quality-management decision makers to separate the most important defects from trivial defects, which helps them to set priorities for needed quality improvement work.

Suppose the number of electric motors being rejected by inspectors for a company has been increasing. Company officials examine the records of several hundred of the motors in which at least one defect was found to determine which defects occurred more frequently. They find that 40% of the defects involved poor wiring, 30% involved a short in the coil, 25% involved a defective plug, and 5% involved cessation of bearings. **Figure 2.10** is a Pareto chart

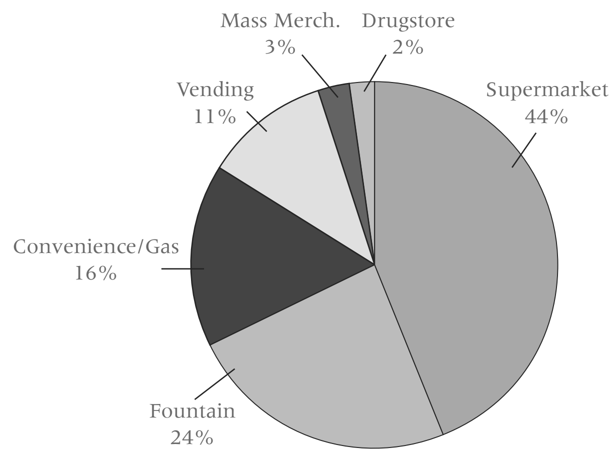
Thinking Critically About Statistics in Business Today

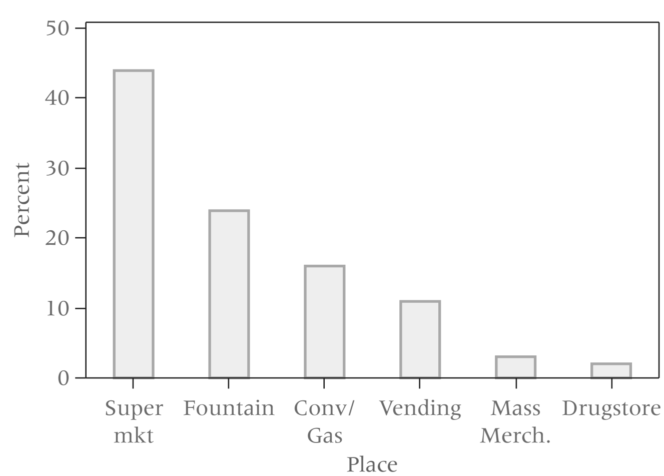
Where Are Soft Drinks Sold?

The soft drink market is an extremely large and growing market in the United States and worldwide. In a recent year, 8.9 billion cases of soft drinks were sold in the United States alone. Where are soft drinks sold? The following data from Sanford C. Bernstein research indicate that the four leading places for soft drink sales are supermarkets, fountains, convenience/gas stores, and vending machines.

PLACE OF SALES	PERCENTAGE
Supermarket	44
Fountain	24
Convenience/gas stations	16
Vending	11
Mass merchandisers	3
Drugstores	2

These data can be displayed graphically several ways. Displayed here are an Excel pie chart and a Minitab bar chart of the data. Some statisticians prefer the histogram or the bar chart over the pie chart because they believe it is easier to compare categories that are similar in size with the histogram or the bar chart rather than the pie chart.





THINGS TO PONDER

1. How might this information be useful to large soft drink companies?
2. How might the packaging of soft drinks differ according to the top four places where soft drinks are sold?
3. How might the distribution of soft drinks differ between the various places where soft drinks are sold?

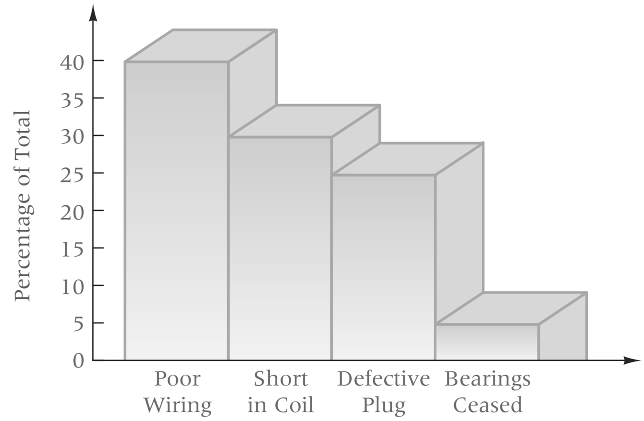
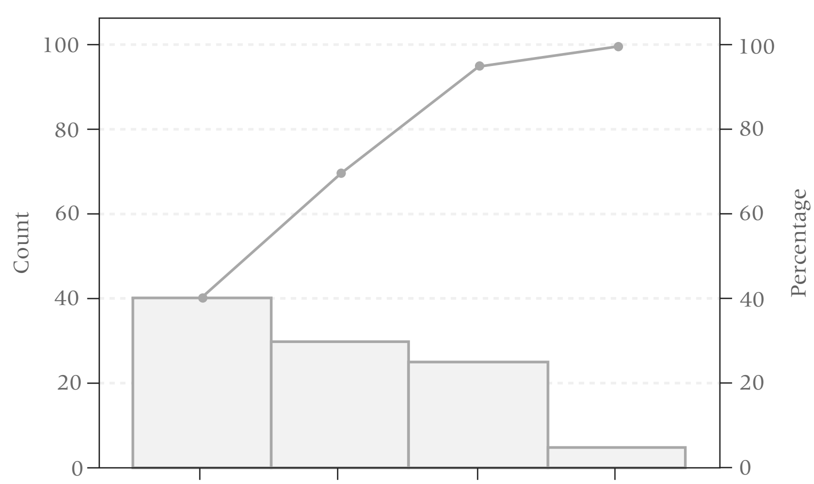


FIGURE 2.10 **Pareto Chart for Electric Motor Problems**

constructed from this information. It shows that the main three problems with defective motors—poor wiring, a short in the coil, and a defective plug—account for 95% of the problems. From the Pareto chart, decision makers can formulate a logical plan for reducing the number of defects.

Company officials and workers would probably begin to improve quality by examining the segments of the production process that involve the wiring. Next, they would study the construction of the coil, then examine the plugs used and the plug-supplier process.

Figure 2.11 is a Minitab rendering of this Pareto chart. In addition to the bar chart analysis, the Minitab Pareto analysis contains a cumulative percentage line graph. Observe the slopes on the line graph. The steepest slopes represent the more frequently occurring problems. As the slopes level off, the problems occur less frequently. The line graph gives the decision maker another tool for determining which problems to solve first.



Type of Defect	Poor Wiring	Short in Coil	Defective Plug	Others
Count	40	30	25	5
Percent	40.0	30.0	25.0	5.0
Cum%	40.0	70.0	95.0	100.0

FIGURE 2.11 **Minitab Pareto Chart for Electric Motor Problems**

2.3 Problems

2.15. Shown here is a list published by Electronics Weekly.com of the top five semiconductor companies in the United States by revenue ($ millions).

FIRM	REVENUE ($ MILLION)
Intel Corp.	51,368
Qualcomm	19,100
Micro + Elpida	16,614
Texas Instruments	12,179
Broadcom	8,360

a. Construct a bar chart to display these data.

b. Construct a pie chart from these data and label the slices with the appropriate percentages.

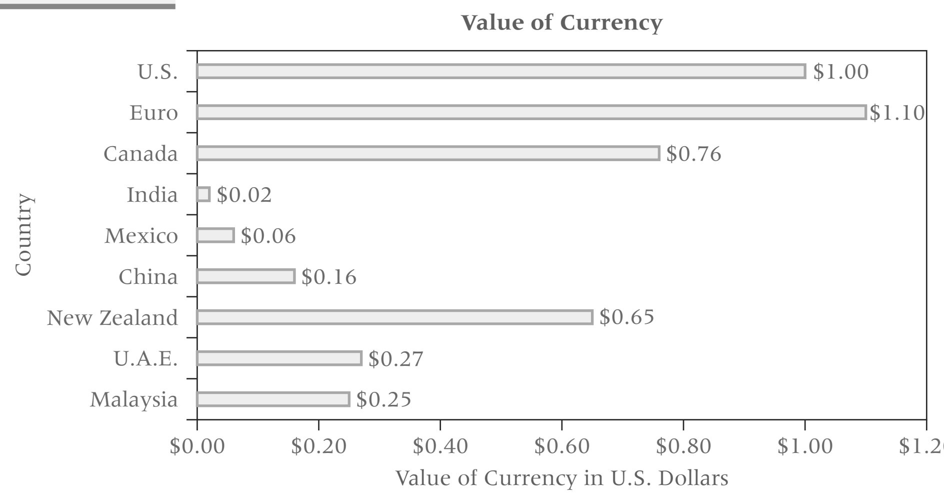
c. Comment on the effectiveness of using a pie chart versus a bar chart to display the revenue of these five companies.

2.16. According to Bureau of Transportation statistics, the largest five U.S. airlines in scheduled system-wide (domestic and international) enplanements in 2013 (passenger numbers in millions) were: Delta with 120.4, Southwest with 115.3, United with 90.1, American with 86.8 and US Airways with 57.0. Construct a pie chart and a bar graph to depict this information.

2.17. The following list shows the top six pharmaceutical companies in the United States by revenue ($ billions) for a recent year as published by Contract Pharma. Use this information to construct a pie chart and a bar graph to represent these six companies and their revenue.

PHARMACEUTICAL COMPANY	SALES
Pfizer	51.2
Merck	40.6
Johnson & Johnson	25.4
Abbott Laboratories	23.1
Eli Lilly	20.6
Bristol-Myers Squibb	17.6

2.18. How do various currencies around the world stack up to the U.S. dollar? Shown below is a bar chart of the value of the currency of various countries in U.S. dollars as of April 2015. The currencies represented here are the Malaysia ringgit, United Arab Emirates dirham, New Zealand dollar, China yuan, Mexico peso, India rupee, Canada dollar, European euro, and U.S. dollar. Study the bar chart and discuss the various currencies as they relate to each other in value and as they compare to the U.S. dollar.



2.19. An airline company uses a central telephone bank and a semi-automated telephone process to take reservations. It has been receiving an unusually high number of customer complaints about its reservation system. The company conducted a survey of customers, asking them whether they had encountered any of the following problems in making reservations: busy signal, disconnection, poor connection, too long a wait to talk to someone, could not get through to an agent, connected with the wrong person. Suppose a survey of 744 complaining customers resulted in the following frequency tally.

NUMBER OF COMPLAINTS	COMPLAINT
184	Too long a wait
10	Transferred to the wrong person
85	Could not get through to an agent
37	Got disconnected
420	Busy signal
8	Poor connection

Construct a Pareto diagram from this information to display the various problems encountered in making reservations.

2.4 Charts and Graphs for Two Variables

It is very common in business statistics to want to analyze two variables simultaneously in an effort to gain insight into a possible relationship between them. For example, business researchers might be interested in the relationship between years of experience and amount of productivity in a manufacturing facility or in the relationship between a person's technology usage and their age. Business statistics has many techniques for exploring such relationships. Two of the more elementary tools for observing the relationships between two variables are cross tabulation and scatter plot.

Cross Tabulation

Cross tabulation is *a process for producing a two-dimensional table that displays the frequency counts for two variables simultaneously*. As an example, suppose a job satisfaction survey of a randomly selected sample of 177 bankers is taken in the banking industry. The bankers are asked how satisfied they are with their job using a 1 to 5 scale where 1 denotes very dissatisfied, 2 denotes dissatisfied, 3 denotes neither satisfied nor dissatisfied, 4 denotes satisfied, and 5 denotes very satisfied. In addition, each banker is asked to report his/her age by using one of the three categories: under 30 years, 30 to 50 years, and over 50 years. Table 2.8 displays how some of the data might look as they are gathered. Note that for each banker the level of their job satisfaction and their age are recorded. By tallying the frequency of responses for each combination of categories between the two variables, the data are cross tabulated according to the two variables. For instance, in this example, there is a tally of how many bankers rated their level of satisfaction as 1 and were under 30 years of age, there is a tally of how many bankers rated their level of satisfaction as 2 and were under 30 years of age, and so on until frequency tallies are determined for each possible combination of the two variables. Table 2.9 shows the completed cross-tabulation table for the banker survey. A cross-tabulation table is sometimes referred to as a contingency table, and Excel refers to such a table as a Pivot Table.

TABLE 2.8 Banker Data Observations by Job Satisfaction and Age

BANKER	LEVEL OF JOB SATISFACTION	AGE
1	4	53
2	3	37
3	1	24
4	2	28
5	4	46
6	5	62
7	3	41
8	3	32
9	4	29
.		
.		
.		
177	3	51

TABLE 2.9 Cross-Tabulation Table of Banker Data

		Age Category			
		Under 30	30–50	Over 50	**Total**
Level of Job Satisfaction	1	7	3	0	10
	2	19	14	3	36
	3	28	17	12	57
	4	11	22	16	49
	5	2	9	14	25
	Total	67	65	45	177

TABLE 2.10 Value of New Construction Over a 35-Year Period

RESIDENTIAL	NONRESIDENTIAL	RESIDENTIAL	NONRESIDENTIAL
169635	96497	252745	169173
155113	115372	228943	167896
149410	96407	197526	135389
175822	129275	232134	120921
162706	140569	249757	122222
134605	145054	274956	127593
195028	131289	251937	139711
231396	155261	281229	153866
234955	178925	280748	166754
266481	163740	297886	177639
267063	160363	315757	175048
263385	164191		

Source: U.S. Census Bureau, *Current Construction Reports* (in millions of constant dollars).

Scatter Plot

A **scatter plot** is a *two-dimensional graph plot of pairs of points from two numerical variables*. The scatter plot is a graphical tool that is often used to examine possible relationships between two variables.

Observe the data in **Table 2.10**. Displayed are the values of new residential and new nonresidential buildings in the United States for various years over a 35-year period. Do these two numerical variables exhibit any relationship? It might seem logical when new construction booms that it would boom in both residential building and in nonresidential building at the same time. However, the Minitab scatter plot of these data displayed in **Figure 2.12** shows somewhat mixed results. The apparent tendency is that more new residential building construction occurs when more new nonresidential building construction is also taking place and less new residential building construction when new nonresidential building construction is also at lower levels. The scatter plot also shows that in some years more new residential building and less new nonresidential building happened at the same time, and vice versa.

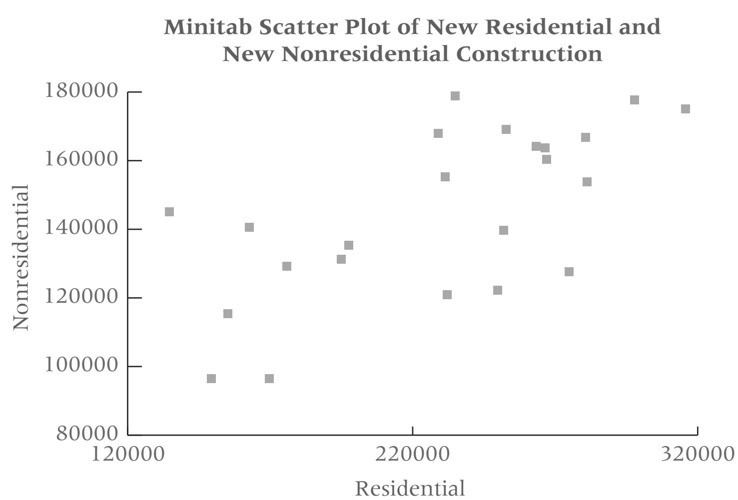


FIGURE 2.12 **Minitab Scatter Plot of New Residential and New Nonresidential Construction**

2.4 Problems

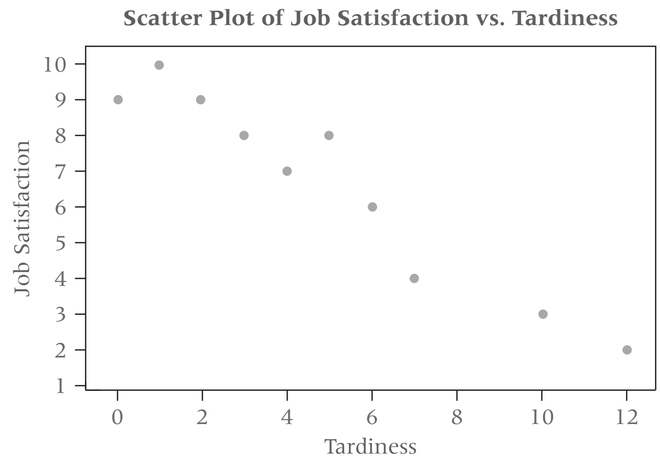
2.20. The U.S. National Oceanic and Atmospheric Administration, National Marine Fisheries Service, publishes data on the quantity and value of domestic fishing in the United States. The quantity (in millions of pounds) of fish caught and used for human food and for industrial products (oil, bait, animal food, etc.) over a decade follows. Is a relationship evident between the quantity used for human food and the quantity used for industrial products for a given year? Construct a scatter plot of the data. Examine the plot and discuss the strength of the relationship of the two variables.

HUMAN FOOD	INDUSTRIAL PRODUCT
3654	2828
3547	2430
3285	3082
3238	3201
3320	3118
3294	2964
3393	2638
3946	2950
4588	2604
6204	2259

2.21. Are the advertising dollars spent by a company related to total sales revenue? The following data represent the advertising dollars and the sales revenues for various companies in a given industry during a recent year. Construct a scatter plot of the data from the two variables and discuss the relationship between the two variables.

ADVERTISING (IN $ MILLIONS)	SALES (IN $ MILLIONS)
4.2	155.7
1.6	87.3
6.3	135.6
2.7	99.0
10.4	168.2
7.1	136.9
5.5	101.4
8.3	158.2

2.22. It seems logical that the number of days per year that an employee is tardy is at least somewhat related to the employee's job satisfaction. Suppose 10 employees are asked to record how satisfied they are with their job on a scale of 0 to 10, with 0 denoting completely unsatisfied and 10 denoting completely satisfied. Suppose also that through human resource records, it is determined how many days each of these employees was tardy last year. The scatter plot below graphs the job satisfaction scores of each employee against the number of days he/she was tardy. What information can you glean from the scatter plot? Does there appear to be any relationship between job satisfaction and tardiness? If so, how might they appear to be related?



2.23. The human resources manager of a large chemical plant was interested in determining what factors might be related to the number of non-vacation days that workers were absent during the past year. One of the factors that the manager considered was the distance that the person commutes to work, wondering if longer commutes resulting in such things as stress on the worker and/or increases in the likelihood of transportation failure might result in more worker absences. The manager studied company records, randomly selected 532 plant workers, and recorded the number of non-vacation days that the workers were absent last year and how far their place of residence is from the plant. The manager then recoded the raw data in categories and created the cross-tabulation table just shown. Study the table and comment on any relationship that may exist between distance to the plant and number of absences.

		One-Way Commute Distance (in miles)		
		0–3	4–10	More than 10
Number of Annual Non-Vacation-Day Absences	0–2	95	184	117
	3–5	21	40	53
	more than 5	3	7	12

2.24. A customer relations expert for a retail tire company is interested in determining if there is any relationship between a customer's level of education and his or her rating of the quality of the tire company's service. The tire company administers a very brief survey to each customer who buys a pair of tires and has them installed at the store. The customer is asked to respond to the quality of the service rendered as either "acceptable" or "unacceptable." In addition, each respondent is asked the level of education attained from the categories of "high school only"or "college degree." These data are gathered on 25 customers and are given below. Use this information to construct a cross-tabulation table. Comment on any relationships that may exist in the table.