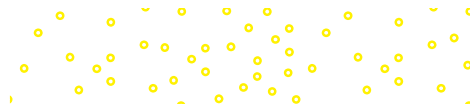




# Essential Statistics

**in Business and Economics**

*Third Edition*



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

### SUPPLY CHAIN MANAGEMENT

Benton

**Purchasing and Supply Chain Management**  
*Third Edition*

Bowersox, Closs, Cooper, and Bowersox  
**Supply Chain Logistics Management**  
*Fifth Edition*

Burt, Petcavage, and Pinkerton  
**Supply Management**  
*Eighth Edition*

Johnson  
**Purchasing and Supply Management**  
*Sixteenth Edition*

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

Stock and Manrodt  
**Supply Chain Management**

### PROJECT MANAGEMENT

Brown and Hyer  
**Managing Projects: A Team-Based Approach**

Larson and Gray  
**Project Management: The Managerial Process**  
*Seventh Edition*

### SERVICE OPERATIONS MANAGEMENT

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

### MANAGEMENT SCIENCE

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

### BUSINESS RESEARCH METHODS

Schindler  
**Business Research Methods**  
*Thirteenth Edition*

### BUSINESS FORECASTING

Keating and Wilson  
**Forecasting and Predictive Analytics**  
*Seventh Edition*

### LINEAR STATISTICS AND REGRESSION

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

### BUSINESS SYSTEMS DYNAMICS

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

### OPERATIONS MANAGEMENT

Cachon and Terwiesch  
**Operations Management**  
*Second Edition*

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

Jacobs and Chase  
**Operations and Supply Chain Management**  
*Fifteenth Edition*

Jacobs and Chase  
**Operations and Supply Chain Management: The Core**  
*Fifth Edition*

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

Stevenson  
**Operations Management**  
*Thirteenth Edition*

Swink, Melnyk, and Hartley  
**Managing Operations Across the Supply Chain**  
*Fourth Edition*

### BUSINESS MATH

Slater and Wittry  
**Practical Business Math Procedures**  
*Thirteenth Edition*

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

### BUSINESS STATISTICS

Bowerman, Drougas, Duckworth, Froelich, Hummel, Moninger, and Schur  
**Business Statistics in Practice**  
*Ninth Edition*

Doane and Seward  
**Applied Statistics in Business and Economics**  
*Sixth Edition*

Doane and Seward  
**Essential Statistics in Business and Economics**  
*Third Edition*

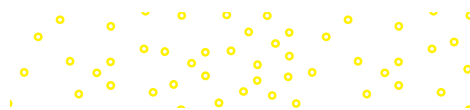
Lind, Marchal, and Wathen  
**Basic Statistics for Business and Economics**  
*Ninth Edition*

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

Jaggia and Kelly  
**Business Statistics: Communicating with Numbers**  
*Third Edition*

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

McGuckian  
**Connect Master: Business Statistics**





# Essential Statistics

**in Business and Economics**

*Third Edition*

**David P. Doane**

*Oakland University*

**Lori E. Seward**

*University of Colorado*

**Mc  
Graw  
Hill**  
Education



## ESSENTIAL STATISTICS IN BUSINESS AND ECONOMICS, THIRD EDITION

Published by McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121. Copyright © 2020 by McGraw-Hill Education. All rights reserved. Printed in the United States of America. Previous editions © 2010, 2008. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written consent of McGraw-Hill Education, including, but not limited to, in any network or other electronic storage or transmission, or broadcast for distance learning.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

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

ISBN 978-1-260-23950-8

MHID 1-260-23950-0

Portfolio Manager: *Noelle Bathurst*

Lead Product Developer: *Michele Janicek*

Product Developer: *Tobi Philips*

Executive Marketing Manager: *Harper Christopher*

Content Project Manager: *Jamie Koch*

Buyer: *Laura Fuller*

Designer: *Matt Diamond*

Content Licensing Specialists: *Ann Marie Jannette*

Cover Image: ©*De Visu/Shutterstock*

Compositor: *SPi Global*

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

### Library of Congress Cataloging-in-Publication Data

Names: Doane, David P., author. | Seward, Lori Welte, 1962- author.

Title: Essential statistics in business and economics / David P. Doane, Oakland University, Lori E. Seward, University of Colorado.

Description: Third edition. | New York, NY : McGraw-Hill Education, [2020]

Identifiers: LCCN 2018028530 | ISBN 9781260239508 (alk. paper)

Subjects: LCSH: Commercial statistics. | Economics—Statistical methods.

Classification: LCC HF1017 .D553 2020 | DDC 519.5—dc23

LC record available at <https://lcn.loc.gov/2018028530>

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

[mheducation.com/highered](http://mheducation.com/highered)

# ABOUT THE AUTHORS



Courtesy of David Doane

## David P. Doane

**David P. Doane** is accredited by the American Statistical Association as a Professional Statistician (PStat<sup>®</sup>). He is professor emeritus in Oakland University's Department of Decision and Information Sciences. He earned his Bachelor of Arts degree in mathematics and economics at the University of Kansas and his PhD from Purdue University's Krannert Graduate School. His research and teaching interests include applied statistics, forecasting, and statistical education. He is co-recipient of three National Science Foundation grants to develop software to teach statistics and to create a computer classroom. He is a longtime member of the American Statistical Association, serving in 2002 as president of the Detroit ASA. He has consulted with government, health care organizations, and local firms. He has published articles in many academic journals. He currently belongs to ASA chapters in San Diego and Orange County/Long Beach.



Courtesy of Lori Seward

## Lori E. Seward

**Lori E. Seward** is a teaching professor in the department of Strategy, Entrepreneurship, and Operations Management in The Leeds School of Business at the University of Colorado in Boulder. She earned her Bachelor of Science and Master of Science degrees in Industrial Engineering at Virginia Tech. After several years working as a reliability and quality engineer in the paper and automotive industries, she earned her PhD from Virginia Tech and joined the faculty at The Leeds School in 1998. Her teaching focuses on developing pedagogy that uses technology to create a collaborative learning environment in large undergraduate and MBA statistics courses. She has been the coordinator of the undergraduate core business statistics course and currently teaches as well as coordinates the core statistics course for both the Leeds full-time and Executive MBA programs. She is also responsible for coordinating the undergraduate program in Operations Management. Dr. Seward is the faculty director of the Leeds Global Seminars in International Operations Management that take place in China. Outside of Leeds, Dr. Seward has served as the chair of the annual INFORMS Teachers' Workshop, has developed courses for international business programs, and has provided statistical consulting in support of the UN's Clean Development Mechanism. Her most recent article, co-authored with David Doane, was published in the *Journal of Statistics Education* (2011).

# DEDICATION

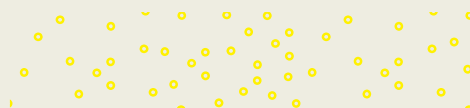
To Robert Hamilton Doane-Solomon

David

To all my students who challenged me to make statistics relevant to their lives.

Lori

▼



# FROM THE

“How often have you heard people/students say about a particular subject, ‘I’ll never use this in the real world’? I thought statistics was a bit on the ‘math-geeky’ side at first. Imagine my horror when I saw  $\alpha$ ,  $R^2$ , and correlations on several financial reports at my current job (an intern position at a financial services company). I realized then that I had better try to understand some of this stuff.”

—Jill Odette (an introductory statistics student)

As recently as a decade ago, our students used to ask us, “**How** do I use statistics?” Today we more often hear, “**Why** should I use statistics?” *Essential Statistics in Business and Economics* has attempted to provide real meaning to the use of statistics in our world by using real business situations and real data and appealing to your need to know *why* rather than just *how*.

With over 50 years of teaching statistics between the two of us, we feel we have something to offer. Seeing how students have changed as the new century unfolds has required us to adapt and seek out better ways of instruction. So we wrote *Essential Statistics in Business and Economics* to meet four distinct objectives.

**Objective 1: Communicate the Meaning of Variation in a Business Context** Variation exists everywhere in the world around us. Successful businesses know how to measure variation. They also know how to tell when variation should be responded to and when it should be left alone. We’ll show how businesses do this.

**Objective 2: Use Realistic Business Applications** We offer examples, case studies, and problems from current research or real applications whenever possible. Hypothetical data are used when it seems the best way to illustrate a concept. You can usually tell the difference by examining the footnotes citing the source.

**Objective 3: Incorporate Current Statistical Practices and Offer Practical Advice** With the increased reliance on computers and data analytics, statistics practitioners have changed the way they use statistical tools. We’ll show the current practices and explain why they are used the way they are. We also will tell you when each technique should *not* be used.

**Objective 4: Provide More In-Depth Explanation of the Why and Let the Software Take Care of the How** It is critical to understand the importance of communicating with data. Today’s technology makes it easier to summarize and display data than ever before. We demonstrate easily mastered techniques with commonly available software. We emphasize the idea that there are risks in decision making and that those risks should be quantified and considered in business decisions.

Our experience tells us that students want to be given credit for the experience they bring to the college classroom. We have tried to honor this by choosing examples and exercises set in situations that will draw on students’ already vast knowledge of the world and skills gained from other classes. We emphasize thinking about data, choosing appropriate analytic tools, using computers effectively, and recognizing limitations of statistics. We give practical advice on handling ethical issues faced by data analysts, spotting barriers to critical thinking, improving communication within teams, and avoiding common “rookie” errors.

## What Is “Essential”?

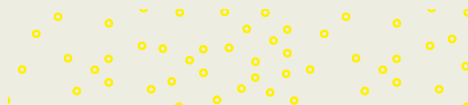
We cannot know which topics will turn out to be essential to each student’s career, so instructors struggle to fit a lot into each class. But there must be trade-offs, particularly in an “essentials” textbook that is designed to be “lean.” For example, here are some contemporary issues that affect textbook length, with valid arguments both *pro* and *con*.

- With Excel available, does the textbook still need full tables (e.g., Poisson, binomial,  $F$ )? Only some of them? Which ones?
- Which statistical distributions can be omitted? Should the text explain approximations (e.g., binomial vs Poisson)?
- Which non-parametric tests (if any) should an “essentials” text cover? What about time series data? Logistic regression?
- How much depth of coverage is “essential” for Bayes’ Theorem? Survey methods? Finite populations? ANOVA? Advanced data analytics tools?

We have made what we hope are appropriate choices and have provided supplemental video tutorials and *LearningStats* demonstrations to extend coverage where more depth is needed.

## What’s New in This Third Edition?

In this edition, we have listened to you and have made many changes that you asked for. We sought advice from students and faculty who are currently using the textbook, reviewers at a variety of colleges and universities, and participants in focus groups on teaching statistics with technology. At the end of this preface is a detailed list of chapter-by-chapter improvements, but here are just a few of them:



# AUTHORS

- Updated Excel support, including screen shots, menus, and functions.
- Introduction to the topic of Analytics and how it fits in with Business Statistics.
- More focus on Excel, moving most screenshots from *MegaStat* and Minitab to end-of-chapter *Software Supplements*.
- Updated exercises with emphasis on compatibility with Connect<sup>®</sup>.
- Updated test bank questions matched with topics and learning objectives.
- Expanded treatment of regression, including multiplicative models, interaction effects, and chapter sections dedicated to logistic regression.
- Rewritten instructor’s manual with step-by-step solutions.
- New and updated Mini Cases for economics and business.
- New and updated exercise data sets, web links, *Big Data Sets*, and *Related Reading*.
- Downloadable supplements from Connect<sup>®</sup> including *LearningStats* demonstrations and video tutorials (both PC and Mac) for Excel, *MegaStat*, Minitab, and JMP.

## Software

Excel is used throughout this book because it is available everywhere. Some calculations are illustrated using *MegaStat*, an Excel add-in whose Excel-based menus and spreadsheet format offer more capability than Excel’s Data Analysis Tools. Minitab menus and examples also are included to point out similarities and differences of these tools. To assist students who need extra help or “catch-up” work, the text website contains tutorials or demonstrations on using Excel, Minitab, or *MegaStat* for the tasks of each chapter. At the end of each chapter is a list of *LearningStats* demonstrations that illustrate the concepts from the chapter. These demonstrations can be found in the Connect product for this text. Short video tutorials (Excel, *MegaStat*, Minitab, JMP) are available to users of Connect<sup>®</sup>.

## Math Level

The assumed level of mathematics is precalculus, though there are rare references to calculus where it might help the better-trained reader. All but the simplest proofs and derivations are omitted. Key assumptions are stated clearly. We advise what to do when these assumptions are not fulfilled. Worked examples are included for basic calculations, but the textbook does assume that computers will do most calculations after the statistics class is taken, so *interpretation* is paramount. End-of-chapter references and suggested websites allow interested readers to deepen their understanding.

## Exercises

Simple practice exercises are placed within each section. End-of-chapter exercises tend to be more integrative or to be embedded in more realistic contexts. Attention has been given to revising exercises so that they have clear-cut answers that are matched to specific learning objectives. A few exercises invite short answers rather than just quoting a formula. Answers to most odd-numbered exercises are in the back of the book (all of the answers are in the instructor’s manual).

## LearningStats

Connect<sup>®</sup> users can access *LearningStats*, a collection of Excel spreadsheets, Word documents, and PowerPoints for each chapter intended to let students explore data and concepts that interest them. *LearningStats* includes explanations on topics such as how to write effective reports, how to perform calculations, or how to make effective charts. It also includes topics that did not appear prominently in the textbook (e.g., partial  $F$  test, Durbin–Watson test, sign test, bootstrap simulation, and logistic regression). Instructors can use *LearningStats* PowerPoint presentations in the classroom and Connect<sup>®</sup> users also can use them for review. No instructor can “cover everything,” but students can be encouraged to explore *LearningStats* demonstrations, perhaps with an instructor’s guidance.

David P. Doane  
Lori E. Seward



# HOW ARE THE CHAPTERS ORGANIZED

## Chapter Contents

Each chapter begins with a short list of section topics that are covered in the chapter.

## Chapter Learning Objectives

Each chapter includes a list of learning objectives students should be able to attain upon reading and studying the chapter material. Learning objectives give students an overview of what is expected and identify the goals for learning. Learning objectives also appear next to chapter topics in the margins.

**CHAPTER CONTENTS**

- 1.1 What Is Statistics?
- 1.2 Why Study Statistics?
- 1.3 Statistics in Business
- 1.4 Statistical Challenges
- 1.5 Critical Thinking

**CHAPTER LEARNING OBJECTIVES**

**LO** When you finish this chapter, you should be able to

- LO 1-1** Define statistics and explain some of its uses.
- LO 1-2** List reasons for a business student to study statistics.
- LO 1-3** Explain the uses of statistics in business.
- LO 1-4** State the common challenges facing business professionals using statistics.
- LO 1-5** List and explain common statistical pitfalls.

## Section Exercises

Multiple section exercises are found throughout the chapter so that students can focus on material just learned.

**SECTION EXERCISES**

**connect**

3.1 (a) Make a stem-and-leaf plot for these 24 observations on the number of customers who used a downtown CitiBank ATM during the noon hour on 24 consecutive workdays. (b) Make a dot plot of the ATM data. (c) Describe these two displays. (*Hint:* Refer to center, variability, and shape.) CitiBank

39	32	21	26	19	27	32	25
18	26	34	18	31	35	21	33
33	9	16	32	35	42	15	24

3.2 (a) Make a stem-and-leaf plot for the number of defects per 100 vehicles for these 32 brands. (b) Make a dot plot of the defects data. (c) Describe these two displays. (*Hint:* Refer to center, variability, and shape.) JDPower

## Mini Cases

Every chapter includes two or three mini cases, which are solved applications. They show and illustrate the analytical application of specific statistical concepts at a deeper level than the examples.

**4.2**

**Mini Case**

**Prices of Lipitor®**

Prescription drug prices vary across the United States and even among pharmacies in the same city. A consumer research group examined prices for a 30-day supply of Lipitor® (a cholesterol-lowering prescription drug) in three U.S. cities at various pharmacies. Attention has recently been focused on prices of such drugs because recent medical research has suggested more aggressive treatment of high cholesterol levels in patients at risk for heart disease. This poses an economic issue for government because Medicare is expected to pay some of the cost of prescription drugs. It is also an issue for Pfizer, the maker of Lipitor®, who expects a fair return on its investments in research and patents. Finally, it is an issue for consumers who seek to shop wisely.

From the dot plots in Figure 4.14, we gain an impression of the *variability* of the data (the *range* of prices for the drug in each city) as well as the *center* of the data (the middle or typical data values). Lipitor® prices vary from about \$60 to about \$91 and typically are in the \$70s. The dot plots suggest that Providence tends to have higher prices, and New Orleans lower prices, though there is considerable variation among pharmacies.

**FIGURE 4.14** Dot Plots for Lipitor® Prices Lipitor

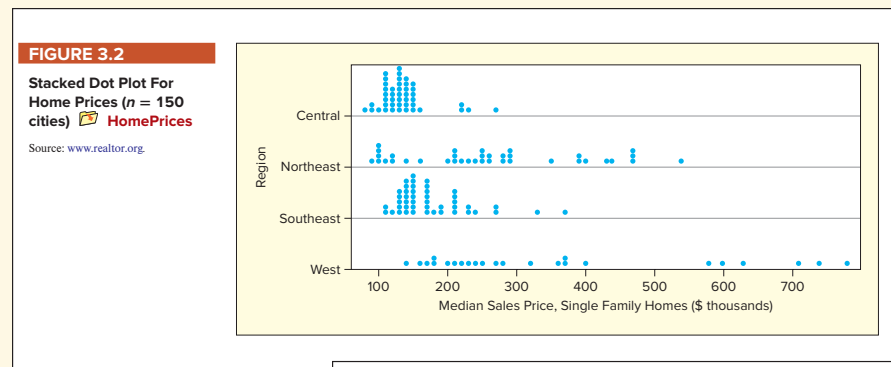
City	Price (\$)
Grand Rapids	65
Grand Rapids	66
Grand Rapids	67
Grand Rapids	68
Grand Rapids	69
Grand Rapids	70
Grand Rapids	71
Grand Rapids	72
Grand Rapids	73
Grand Rapids	74
Grand Rapids	75
Grand Rapids	76
Grand Rapids	77
Grand Rapids	78
Grand Rapids	79
Grand Rapids	80
Grand Rapids	81
Grand Rapids	82
Grand Rapids	83
Grand Rapids	84
Grand Rapids	85
Providence	70
Providence	71
Providence	72
Providence	73
Providence	74
Providence	75
Providence	76
Providence	77
Providence	78
Providence	79
Providence	80
Providence	81
Providence	82
Providence	83
Providence	84
Providence	85
Providence	86
Providence	87
Providence	88
Providence	89
Providence	90
Providence	91
New Orleans	60
New Orleans	61
New Orleans	62
New Orleans	63
New Orleans	64
New Orleans	65
New Orleans	66
New Orleans	67
New Orleans	68
New Orleans	69
New Orleans	70
New Orleans	71
New Orleans	72
New Orleans	73
New Orleans	74
New Orleans	75
New Orleans	76
New Orleans	77
New Orleans	78
New Orleans	79
New Orleans	80
New Orleans	81
New Orleans	82
New Orleans	83
New Orleans	84
New Orleans	85
New Orleans	86
New Orleans	87
New Orleans	88
New Orleans	89
New Orleans	90
New Orleans	91



# TO PROMOTE STUDENT LEARNING?

## Figures and Tables

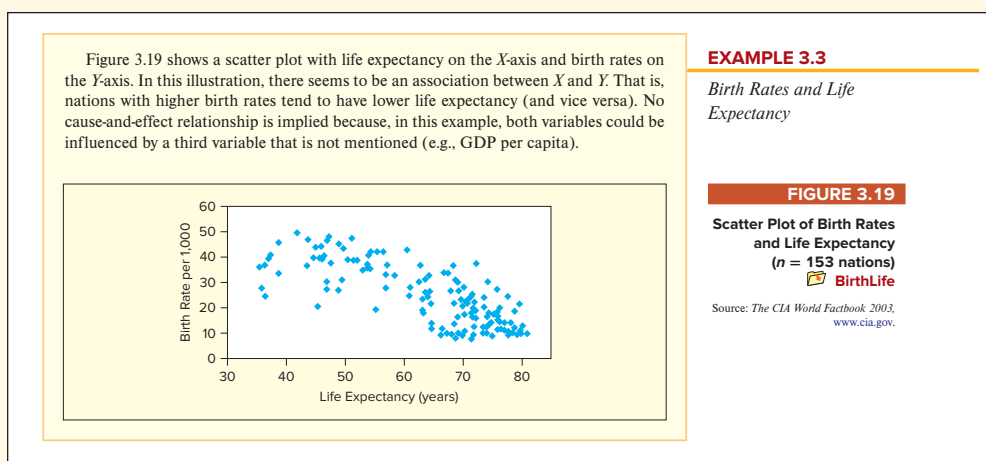
Throughout the text, there are hundreds of charts, graphs, tables, and spreadsheets to illustrate statistical concepts being applied. These visuals help stimulate student interest and clarify the text explanations.



		<b>TABLE 2.5</b>
		<b>Random Sampling Methods</b>
Simple random sample	Use random numbers to select items from a list (e.g., Visa cardholders).	
Systematic sample	Select every $k$ th item from a list or sequence (e.g., restaurant customers).	
Stratified sample	Select randomly within defined strata (e.g., by age, occupation, gender).	
Cluster sample	Select random geographical regions (e.g., zip codes) that represent the population.	

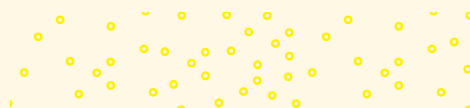
## Examples

Examples of interest to students are taken from published research or real applications to illustrate the statistics concept. For the most part, examples are focused on business, but there are also some that are more general and don't require any prerequisite knowledge. And there are some that are based on student projects.



## Data Set Icon

A data set icon is used throughout the text to identify data sets used in the figures, examples, and exercises that are included in Connect for the text.



# HOW DOES THIS TEXT REINFORCE

## Chapter Summary

Chapter summaries provide an overview of the material covered in the chapter.

### CHAPTER SUMMARY

The **mean** and **median** describe a sample's **center** and also indicate **skewness**. The **mode** is useful for discrete data with a small range. The **trimmed mean** eliminates extreme values. The **geometric mean** mitigates high extremes but cannot be used when zeros or negative values are present. The **midrange** is easy to calculate but is sensitive to extremes. Variability is typically measured by the **standard deviation** while relative dispersion is given by the **coefficient of variation** for nonnegative data. **Standardized data** reveal **outliers** or unusual data values, and the **Empirical Rule** offers a comparison with a normal distribution. In measuring dispersion, the **mean absolute deviation** or **MAD** is easy to understand but lacks nice mathematical properties. **Quartiles** are meaningful even for fairly small data sets, while **percentiles** are used only for large data sets. **Box plots** show the quartiles and data range. The **correlation coefficient** measures the degree of linearity between two variables. The **covariance** measures the degree to which two variables move together. We can estimate many common descriptive statistics from **grouped data**. Sample coefficients of **skewness** and **kurtosis** allow more precise inferences about the **shape** of the population being sampled instead of relying on histograms.

## Key Terms

Key terms are highlighted and defined within the text. They are also listed at the ends of chapters to aid in reviewing.

### KEY TERMS

Center	Variability	Shape	Other
geometric mean	Chebyshev's Theorem	bimodal distribution	box plot
mean	coefficient of variation	kurtosis	covariance
median	Empirical Rule	kurtosis coefficient	five-number summary
midhinge	mean absolute deviation	leptokurtic	interquartile range
midrange	outliers	mesokurtic	method of medians
mode	population variance	multimodal	quartiles
trimmed mean	range	distribution	sample correlation coefficient
weighted mean	sample variance	negatively skewed	
	standard deviation	Pearson 2 skewness coefficient	
	standardized data	platykurtic	
	two-sum formula	positively skewed	
	z-score	Schild's Rule	
		skewed left	
		skewed right	
		skewness	
		skewness coefficient	
		symmetric data	

## Commonly Used Formulas

Some chapters provide a listing of commonly used formulas for the topic under discussion.

### Commonly Used Formulas in Descriptive Statistics

Sample mean:	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Geometric mean:	$G = \sqrt[n]{x_1 x_2 \dots x_n}$
Growth rate:	$GR = \sqrt[n-1]{\frac{x_n}{x_1}} - 1$
Range:	Range = $x_{\max} - x_{\min}$
Midrange:	Midrange = $\frac{x_{\max} + x_{\min}}{2}$
Sample standard deviation:	$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$

## Chapter Review

Each chapter has a list of questions for student self-review or for discussion.

### CHAPTER REVIEW




1. What are descriptive statistics? How do they differ from visual displays of data?
2. Explain each concept: (a) center, (b) variability, and (c) shape.
3. (a) Why is sorting usually the first step in data analysis? (b) Why is it useful to begin a data analysis by thinking about how the data were collected?
4. List strengths and weaknesses of each measure of center and write its Excel function: (a) mean, (b) median, and (c) mode.
5. (a) Why must the deviations around the mean sum to zero? (b) What is the position of the median in the data array when  $n$  is even? When  $n$  is odd? (c) Why is the mode of little use in continuous data? (d) For what type of data is the mode most useful?
6. (a) What is a bimodal distribution? (b) Explain two ways to detect skewness.

# STUDENT LEARNING?

## Chapter Exercises

Exercises give students an opportunity to test their understanding of the chapter material. Exercises are included at the ends of sections and at the ends of chapters. Some exercises contain data sets, identified by data set icons. Data sets can be accessed through Connect and used to solve problems in the text.

### EXCEL PROJECTS

- 4.87 (a) Use Excel functions to calculate the mean and standard deviation for weekend occupancy rates (percent) in nine resort hotels during the off-season. (b) What conclusion would a casual observer draw about center and variability, based on your statistics? (c) Now calculate the median for each sample. (d) Make a dot plot for each sample. (e) What did you learn from the medians and dot plots that was not apparent from the means and standard deviations?  **Occupancy**

Observation	Week 1	Week 2	Week 3	Week 4
1	32	33	38	37
2	41	35	39	42
3	44	45	39	45
4	47	50	40	46
5	50	52	56	47
6	53	54	57	48
7	56	58	58	50
8	59	59	61	67
9	68	64	62	68












## More Learning Resources

*LearningStats* provides a means for Connect users to explore data and concepts at their own pace. Applications that relate to the material in the chapter are identified by topic at the end of each chapter.

### CHAPTER 4 More Learning Resources

You can access these *LearningStats* demonstrations through McGraw-Hill's Connect® to help you understand descriptive statistics.



Topic	LearningStats Demonstrations
Overview	<ul style="list-style-type: none"> <li> Describing Data</li> <li> Using MegaStat</li> <li> Using MINITAB</li> </ul>
Descriptive statistics	<ul style="list-style-type: none"> <li> Basic Statistics</li> <li> Quartiles</li> <li> Box Plots</li> <li> Grouped Data</li> <li> Significant Digits</li> </ul>
ScreenCam Tutorials	<ul style="list-style-type: none"> <li> Using MegaStat</li> <li> Excel Descriptive Statistics</li> <li> Excel Scatter Plots</li> </ul>

Key:  = PowerPoint  = Excel  = PDF  = ScreenCam Tutorials

## Exam Review Questions

At the end of a group of chapters, students can review the material they covered in those chapters. This provides them with an opportunity to test themselves on their grasp of the material.

- Which type of statistic (descriptive, inferential) is each of the following?
  - Estimating the default rate on all U.S. mortgages from a random sample of 500 loans.
  - Reporting the percent of students in your statistics class who use Verizon.
  - Using a sample of 50 iPhones to predict the average battery life in typical usage.
- Which is *not* an ethical obligation of a statistician? Explain.
  - To know and follow accepted procedures.
  - To ensure data integrity and accurate calculations.
  - To support client wishes in drawing conclusions from the data.
- "Driving without a seat belt is not risky. I've done it for 25 years without an accident." This *best* illustrates which fallacy?
  - Unconscious bias.
  - Conclusion from a small sample.
  - Post hoc* reasoning.
- Which data type (categorical, numerical) is each of the following?
  - Your current credit card balance.
  - Your college major.
  - Your car's odometer mileage reading today.
- Give the type of measurement (nominal, ordinal, interval, ratio) for each variable.
  - Length of time required for a randomly chosen vehicle to cross a toll bridge.
  - Student's ranking of five cell phone service providers.
  - The type of charge card used by a customer (Visa, MasterCard, AmEx, Other).

### EXAM REVIEW QUESTIONS FOR CHAPTERS 1–4



# connect®

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

## SUCCESSFUL SEMESTERS INCLUDE CONNECT

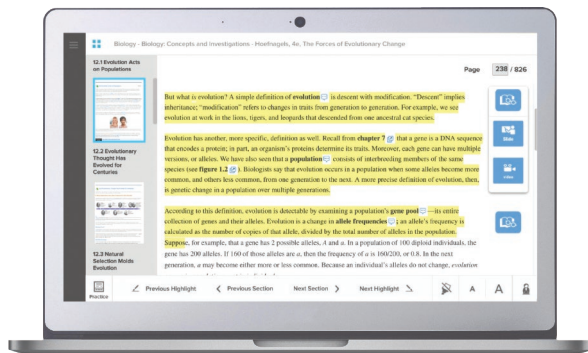
### FOR INSTRUCTORS

#### You're in the driver's seat.

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

# 65%

Less Time  
Grading

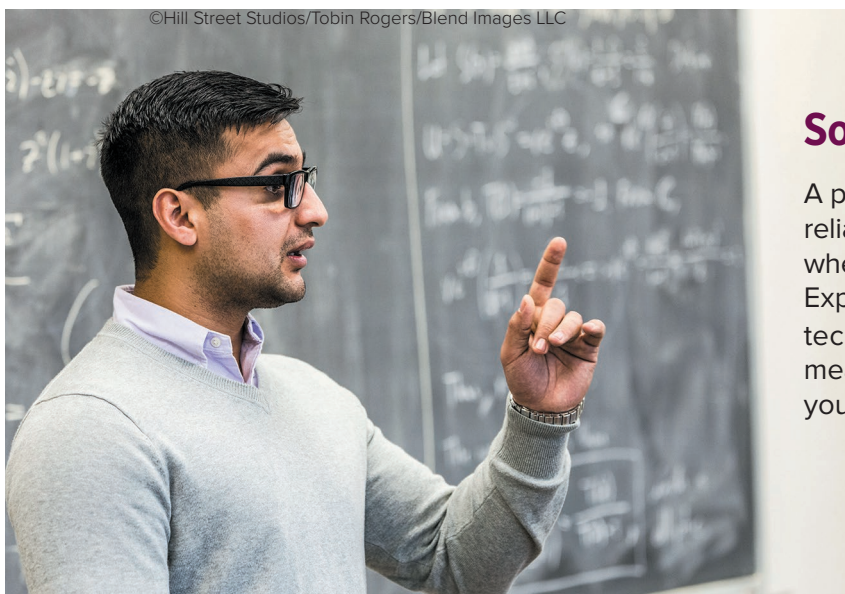
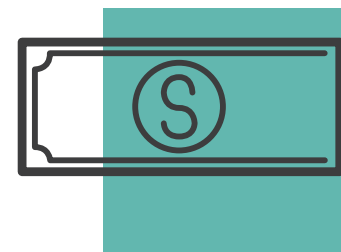


#### They'll thank you for it.

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

#### Make it simple, make it affordable.

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



#### Solutions for your challenges.

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

## FOR STUDENTS

### Effective, efficient studying.

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



©Shutterstock/wavebreakmedia

“I really liked this app—it made it easy to study when you don't have your textbook in front of you.”

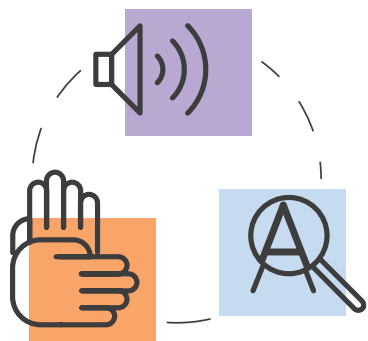
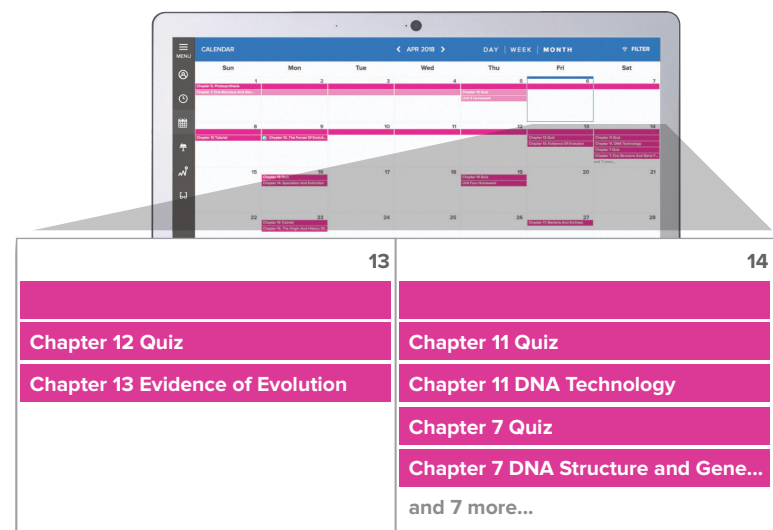
- Jordan Cunningham,  
Eastern Washington University

### Study anytime, anywhere.

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

### No surprises.

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



### Learning for everyone.

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



# ADDITIONAL CONNECT FEATURES

**Excel Data Sets** A convenient feature is the inclusion of an Excel data file link in many problems using data files in their calculation. The link allows students to easily launch into Excel, work the problem, and return to Connect to key in the answer.

Chapter Exercise 5-92  
 High levels of cockpit noise in an aircraft can damage the hearing of pilots who are exposed to this hazard for many hours. Cockpit noise in a jet aircraft is mostly due to airflow at hundreds of miles per hour. This 3x3 contingency table shows 61 observations of data collected by an airline pilot using a handheld sound meter in a Boeing 727 cockpit. Noise level is defined as "low" (under 88 decibels), "medium" (88 to 91 decibels), or "high" (92 decibels or more). There are three flight phases (climb, cruise, descent).

Noise Level	Flight Phase			Row Total
	Climb (B)	Cruise (C)	Descent (D)	
Low (L)	6	2	6	14
Medium (M)	18	3	8	29
High (H)	1	3	14	18
Column Total	25	8	28	61

[Click here for the Excel Data File](#)

(a) Calculate the following probabilities: (Round your answers to 4 decimal places.)

i.  $P(B)$

ii.  $P(L)$

iii.  $P(H | C)$

**Guided Examples** These narrated video walkthroughs provide students with step-by-step guidelines for solving selected exercises similar to those contained in the text. The student is given personalized instruction on how to solve a problem by applying the concepts presented in the chapter. The narrated voiceover shows the steps to take to work through an exercise. Students can go through each example multiple times if needed.

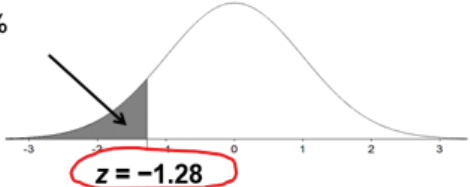
## Chapter 7

### Find Z scores associated with Standard Normal Areas

Find the associated z-score for each of the following standard normal areas using Appendix C-2 or Excel 2010

a. Lowest 10%  
 b. Middle 80%

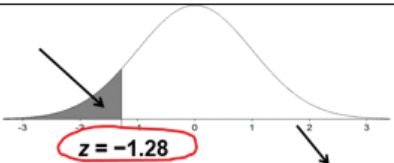
a. Lowest 10%



**z = -1.28**

Excel:  $z = \text{NORM.S.INV}(0.10) = -1.28155$   
 Or rounded to 2 decimal places  $z = -1.28$

a. Lowest 10%



**z = -1.28**

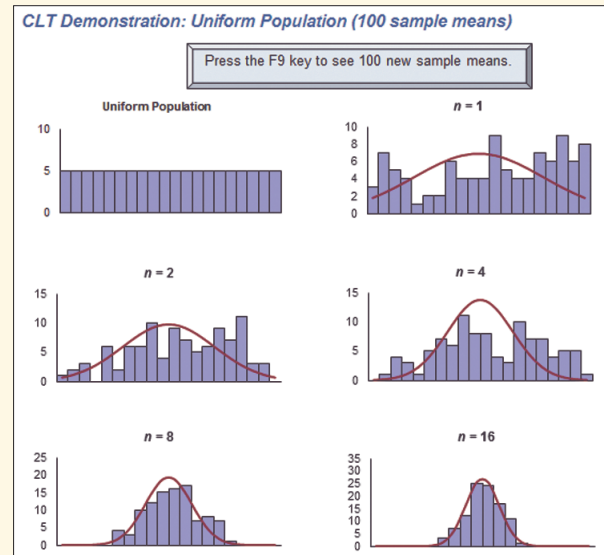
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.7	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008
-3.6	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013	0.0013	0.0012	0.0012	0.0011
-3.5	0.0023	0.0022	0.0022	0.0021	0.0020	0.0019	0.0019	0.0018	0.0017	0.0017
-3.4	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025	0.0024	0.0023
-3.3	0.0044	0.0043	0.0042	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036	0.0035
-3.2	0.0059	0.0058	0.0057	0.0056	0.0055	0.0054	0.0053	0.0052	0.0051	0.0050
-3.1	0.0078	0.0077	0.0076	0.0075	0.0074	0.0073	0.0072	0.0071	0.0070	0.0069
-3.0	0.0107	0.0106	0.0105	0.0104	0.0103	0.0102	0.0101	0.0100	0.0099	0.0098
-2.9	0.0149	0.0148	0.0147	0.0146	0.0145	0.0144	0.0143	0.0142	0.0141	0.0140
-2.8	0.0207	0.0206	0.0205	0.0204	0.0203	0.0202	0.0201	0.0200	0.0199	0.0198
-2.7	0.0287	0.0286	0.0285	0.0284	0.0283	0.0282	0.0281	0.0280	0.0279	0.0278
-2.6	0.0398	0.0397	0.0396	0.0395	0.0394	0.0393	0.0392	0.0391	0.0390	0.0389
-2.5	0.0540	0.0539	0.0538	0.0537	0.0536	0.0535	0.0534	0.0533	0.0532	0.0531
-2.4	0.0715	0.0714	0.0713	0.0712	0.0711	0.0710	0.0709	0.0708	0.0707	0.0706
-2.3	0.0938	0.0937	0.0936	0.0935	0.0934	0.0933	0.0932	0.0931	0.0930	0.0929
-2.2	0.1255	0.1254	0.1253	0.1252	0.1251	0.1250	0.1249	0.1248	0.1247	0.1246
-2.1	0.1677	0.1676	0.1675	0.1674	0.1673	0.1672	0.1671	0.1670	0.1669	0.1668
-2.0	0.2243	0.2242	0.2241	0.2240	0.2239	0.2238	0.2237	0.2236	0.2235	0.2234
-1.9	0.2981	0.2980	0.2979	0.2978	0.2977	0.2976	0.2975	0.2974	0.2973	0.2972
-1.8	0.3944	0.3943	0.3942	0.3941	0.3940	0.3939	0.3938	0.3937	0.3936	0.3935
-1.7	0.5120	0.5119	0.5118	0.5117	0.5116	0.5115	0.5114	0.5113	0.5112	0.5111
-1.6	0.6552	0.6551	0.6550	0.6549	0.6548	0.6547	0.6546	0.6545	0.6544	0.6543
-1.5	0.8340	0.8339	0.8338	0.8337	0.8336	0.8335	0.8334	0.8333	0.8332	0.8331
-1.4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

# WHAT RESOURCES ARE AVAILABLE FOR STUDENTS?

The following software tools are available to assist students in understanding concepts and solving problems.

## LearningStats

*LearningStats* allows students to explore data and concepts at their own pace. It includes demonstrations, simulations, and tutorials that can be downloaded from Connect.



## MegaStat® for Excel®

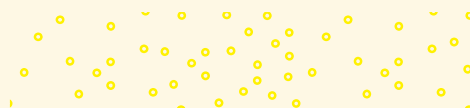
Access Card (ISBN: 0077426274) or online purchase at [www.mhhe.com/megastat](http://www.mhhe.com/megastat).

*MegaStat* is a full-featured Excel add-in that is available to be packaged with this text. It performs statistical analyses within an Excel workbook. It does basic functions such as descriptive statistics, frequency distributions, and probability calculations as well as hypothesis testing, ANOVA, and regression.

*MegaStat* output is carefully formatted, and ease-of-use features include Auto Expand for quick data selection and Auto Label detect. Because *MegaStat* is easy to use, students can focus on learning statistics without being distracted by the software. *MegaStat* is always available from Excel's main menu. Selecting a menu item pops up a dialog box. *MegaStat* is updated continuously to work with the latest versions of Excel for Windows and Macintosh users.

## Minitab®

Free trials and academic versions are available from Minitab at [www.minitab.com](http://www.minitab.com).



# WHAT RESOURCES ARE AVAILABLE FOR INSTRUCTORS?

Instructor resources are available through the Connect course at [connect.mheducation.com](http://connect.mheducation.com). Resources include a complete Instructor's Manual in Word format, the complete Test Bank in both Word files and computerized TestGen format, Instructor PowerPoint slides, text art files, and more.

## TestGen

TestGen is a complete, state-of-the-art test generator and editing application software that allows instructors to quickly and easily select test items from McGraw Hill's TestGen test bank content and to organize, edit, and customize the questions and answers to rapidly generate paper tests. Questions can include stylized text, symbols, graphics, and equations that are inserted directly into questions using built-in mathematical templates. TestGen's random generator provides the option to display different text or calculated number values each time questions are used. With both quick-and-simple test creation and flexible and robust editing tools, TestGen is a test generator system for today's educators.

## Online Course Management

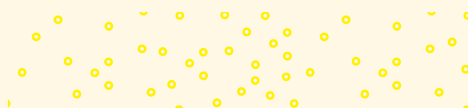
No matter what online course management system you use (WebCT, Blackboard, or eCollege), we have a course content ePack available for your course. Our new ePacks are specifically designed to make it easy for students to navigate and access content online. For help, our online Digital Learning Consultants are ready to assist you with your online course needs. They provide training and will answer any questions you have throughout the life of your adoption.

McGraw-Hill Higher Education and Blackboard have teamed up. What does this mean for you?

1. **Single sign-on.** Now you and your students can access McGraw-Hill's Connect™ and Create™ right from within your Blackboard course—all with one single sign-on.
2. **Integration of content and tools.** You get a single sign-on with Connect™ and Create™, and you also get integration of McGraw-Hill content and content engines right into Blackboard. Whether you're choosing a book for your course or building Connect™ assignments, all the tools you need are right where you want them—inside of Blackboard.
3. **One grade book.** Keeping several grade books and manually synchronizing grades into Blackboard is no longer necessary. When a student completes an integrated Connect™ assignment, the grade for that assignment automatically (and instantly) feeds your Blackboard grade center.
4. **A solution for everyone.** Whether your institution is already using Blackboard or you just want to try Blackboard on your own, we have a solution for you. McGraw-Hill and Blackboard can now offer you easy access to industry-leading technology and content, whether your campus hosts it, or we do. Be sure to ask your local McGraw-Hill representative for details.

## McGraw-Hill Customer Experience Information

For Customer Support, call **800-331-5094** or visit [www.mhhe.com/support](http://www.mhhe.com/support). One of our customer experience team members will be able to assist you in a timely fashion.





# ACKNOWLEDGMENTS

The authors would like to acknowledge some of the many people who have helped with this book. Thomas W. Lauer and Floyd G. Willoughby permitted quotation of a case study. Morgan Elliott, Karl Majeske, Robin McCutcheon, Kevin Murphy, John Sase, T. J. Wharton, and Kenneth M. York permitted questionnaires to be administered in their classes. Mark Isken, Ron Tracy, and Robert Kushler gave generously of their time as expert statistical consultants. Jonathan G. Koomey of Lawrence Berkeley National Laboratory offered valuable suggestions on visual data presentation.

We are grateful to Farrukh Abbas for his careful scrutiny of the text and for offering ideas on improving the text and exercises. Mark Isken has reliably provided Excel expertise and has suggested health care applications for examples and case studies. The Michigan State Employees Credit Union provided ATM data. The Siena Research Institute has made its poll results available. J.D. Power and Associates generously provided permission to use vehicle quality data. The Public Interest Research Group in Michigan (PIRGIM) has generously shared data from its field survey of prescription drug prices.

Phil Rogers has offered numerous suggestions for improvement in both the textbook exercises and Connect. Milo A. Schield shared his research on “quick rules” for measuring skewness from summarized data. We owe special thanks to Aaron Kennedy and Dave Boennighausen of Noodles & Company; to Mark Gasta, Anja Wallace, and Clifton Pacaro of Vail Resorts; to Jim Curtin and Gordon Backman of Ball Corporation; and to Santosh Lakhan from The Verdeo Group for providing suggestions and access to data for Mini Cases and examples. For reviewing the material on quality, we wish to thank Kay Beau regard of William Beaumont Hospital, and Ellen Barnes and Karry Roberts of Ford Motor Company. Amy Sheikh provided a new Facebook Friends data set, along with other excellent suggestions and reports from the “front lines” of her classes.

A special debt of gratitude is due to Michele Janicek for her direction and support and Tobi Philips for coordinating the project. Thanks to Lloyd Jasingh, Morehead State University, for updating the PowerPoint slides. Special thanks to our accuracy checker: Kevin Schaub, University of Colorado. Thanks to the many reviewers who provided such valuable feedback including criticism that made the book better, some of whom reviewed several drafts of the manuscript. Any remaining errors or omissions are the authors’ responsibility. Thanks, too, to the participants in our focus groups and symposia on teaching business statistics, who have provided teaching ideas and insights from their experiences with students in diverse contexts. We hope you will be able to see in our book and the teaching package consideration of those ideas and insights.

Farrukh Abbas, *Barani Institute of Information Technology (Pakistan)*

Heather Adams, *University of Colorado—Boulder*

Sung Ahn, *Washington State University*

Mostafa Aminzadeh, *Towson University*

Scott Bailey, *Troy University*

Hope Baker, *Kennesaw State University*

Saad Taha Bakir, *Alabama State University*

Adam Bohr, *University of Colorado—Boulder*

Katherine Broneck, *Pima Community College—Downtown*

Mary Beth Camp, *Indiana University—Bloomington*

Alan Cannon, *University of Texas—Arlington*

Deborah Carter, *Coahoma Community College*

Kevin Caskey, *SUNY—New Paltz*

Michael Cervetti, *University of Memphis*

Paven Chennamaneni, *University of Wisconsin—Whitewater*

Alan Chesen, *Wright State University*

Wen-Chyuan Chiang, *University of Tulsa*

Chia-Shin Chung, *Cleveland State University*

Joseph Coleman, *Wright State University—Dayton*

Robert Cutshall, *Texas A&M University—Corpus Christi*

Terry Dalton, *University of Denver*

Douglas Dotterweich, *East Tennessee State University*

Jerry Dunn, *Southwestern Oklahoma State University*

Michael Easley, *University of New Orleans*

Jerry Engeholm, *University of South Carolina—Aiken*

Mark Farber, *University of Miami*

Soheila Kahkashan Fardanesh, *Towson University*

Mark Ferris, *St. Louis University*

Stergios Fotopoulos, *Washington State University*

Vickie Fry, *Westmoreland County Community College*

Joseph Fuhr, *Widener University*

Bob Gillette, *University of Kentucky*

Don Gren, *Salt Lake City Community College*

Karina Hauser, *University of Colorado—Boulder*

Clifford Hawley, *West Virginia University*

Yijun He, *Washington State University*

Natalie Hegwood, *Sam Houston State University*

Allen Humbolt, *University of Tulsa*

Patricia Igo, *Northeastern University*

Alam M. Imam, *University of Northern Iowa*

Marc Isaacson, *Augsburg College*

Kishen Iyengar, *University of Colorado—Boulder*

Christopher Johnson, *University of North Florida*

Jerzy Kamburowski, *University of Toledo*

Bob Kitahara, *Troy University*

Drew Koch, *James Madison University*

Agnieszka Kwapisz, *Montana State University*

Kenneth Lawrence, *New Jersey Institute of Technology*

Bob Lynch, *University of Northern Colorado*

Bradley McDonald, *Northern Illinois University*

Richard McGowan, *Boston College*

Kelly McKillop, *University of Massachusetts*

Larry McRae, *Appalachian State University*

Robert Mee, *University of Tennessee—Knoxville*

John Miller, *Sam Houston State University*

Shelly Moore, *College of Western Idaho*

James E. Moran Jr., *Oregon State University*

Geraldine Moultime, *Northwood University*

Adam Munson, *University of Florida*

Joshua Naranjo, *Western Michigan University*

Anthony Narsing, *Macon State College*

Pin Ng, *Northern Arizona University*

Thomas Obremski, *University of Denver*

Mohammad Reza Oskoorouchi, *California State University—San Marcos*

Ceyhun Ozgur, *Valparaiso University*

Nitin Paranjpe, *Oakland University*

Mahour Mellat Parast, *University of North Carolina—Pembroke*

Eddy Patuwo, *Kent State University*

John Pickett, *University of Arkansas—Little Rock*

James Pokorski, *Virginia Polytechnic Institute & State University*

Stephan Pollard, *California State University—Los Angeles*

Claudia Pragman, *Minnesota State University*

Tammy Prater, *Alabama State University*

Michael Racer, *University of Memphis*

Azar Raiszadeh, *Chattanooga State Community College*

Phil Rogers, *University of Southern California*

Milo A. Schield, *Augsburg College*

Sue Schou, *Idaho State University*

Sankara N. Sethuraman, *Augusta State University*

Don Sexton, *Columbia University*

Thomas R. Sexton, *Stony Brook University*

Murali Shanker, *Kent State University*

Gary W. Smith, *Florida State University*

Courtenay Stone, *Ball State University*

Paul Swanson, *Illinois Central College*

Rahmat Tavallali, *Walsh University*

Deborah Tesch, *Xavier University*

Dharma S. Thiruvaiyaru, *Augusta State University*

Bhavneet Walia, *Western Illinois University*

Jesus M. Valencia, *Slippery Rock University*

Rachel Webb, *Portland State University*

Simone A. Wegge, *City University of New York*

Chao Wen, *Eastern Illinois University*

Alan Wheeler, *University of Missouri—St. Louis*

Blake Whitten, *University of Iowa*

Charles Wilf, *Duquesne University*

Anne Williams, *Gateway Community College*

Janet Wolcott, *Wichita State University*

Frank Xie, *University of South Carolina—Aiken*

Ye Zhang, *Indiana University—Purdue University Indianapolis*

Mustafa R. Yilmaz, *Northeastern University*

# ENHANCEMENTS FOR

Many changes were motivated by advice from reviewers and users of the textbook. Besides hundreds of small edits, these changes were common to most chapters:

- New end-of-chapter *Software Supplements* (MegaStat, Minitab) to allow more focus on Excel within chapters.
- Closer exercise compatibility with *Connect*, *SmartBook*, and *LearnSmart*.
- Updated *Related Readings* and *Web Sources* for students who want to “dive deeper.”
- Revised *LearningStats* demonstrations to illustrate concepts beyond what is possible in a textbook (e.g., simulations).
- Updated test bank (with more feedback) and updated/expanded *Big Data Sets*.
- Improved illustrations, figures, and tables.

## Chapter 1—Overview of Statistics

New Mini Cases (e.g., analytics in business, predicting airfares, GM ignition switches).

More discussion of using statistics in business, working in teams, and jobs for data scientists.

Leaner discussion of critical thinking and a new exercise on critical thinking.

Updated *Related Reading* references.

## Chapter 2—Data Collection

Reorganized learning objectives to give more focus on testable topics.

Improved discussion of binning in frequency distributions

Revised treatment of variables, data types, and measurement levels.

Reorganized presentation of samples, populations, and sampling methods,

New Mini Cases (e.g., Super Bowl audiences).

New, revised, and updated exercises (e.g., housing starts, lightning deaths).

Revised explanation of data collection methods, sources of error.

New discussion of reliability, validity, and survey software.

Updated *Web Data Sources*, *Related Reading*, and *LearningStats* demos.

## Chapter 3—Describing Data Visually

More efficient treatment of key topics and examples.

Updated screenshots and advice for Excel charts, histograms, pivot tables, scatter plots.

Moved *MegaStat* and Minitab screenshots to end-of-chapter *Software Supplement*.

New, revised, and updated exercises (e.g., stock prices, web browsers, TV sales).

Updated *Related Reading* references.

## Chapter 4—Descriptive Statistics

Streamlined discussion of main concepts.

Updated Excel screenshots for descriptive statistics.

Moved *MegaStat* and Minitab screenshots to end-of-chapter *Software Supplement*.

New, revised, and updated exercises (e.g., asset turnover ratios, stock prices, skewness, kurtosis,

consumer expenditure,  $z$ -scores, quartiles, grouped data) and many revised data sets.

New and updated Mini Cases (e.g., U.S. presidents’ ages, car defects over time).

Reorganized and expanded section on covariance and correlation.

A new statistic for measuring skewness when only summarized data are available.

New decision diagram to guide student choice of statistics and graphs.

## Chapter 5—Probability

Revised example of defining compound events.

Revised Mini Cases (e.g., women-owned companies, Bayes Theorem).

New, updated, and revised exercises (e.g., free eBay shipping, YouTube videos, online sales, credit card use, flight delays).

## Chapter 6—Discrete Probability Distributions

Reorganized learning objectives to give more focus on testable topics.

Improved topic placement on how to recognize each type of distribution.

Updated Excel screenshots and menus.

New, revised, and updated exercises (e.g., music festival tickets, inner tube rentals).

## Chapter 7—Continuous Probability Distributions

Reorganized learning objectives to give more control by testable topic.

Revised discussion of expected value and variance.

New, revised, and updated exercises (e.g., bus arrivals, heart rates, power surges, defect rates, expected value).

New exercises on using Excel functions.

Updated Excel screenshots and instructions.

New illustration of exponential distribution families and middle areas.

## Chapter 8—Sampling Distributions and Estimation

Reorganized learning objectives to give more focus on testable topics.

Major rewrite of sections on Central Limit Theorem, sampling error, estimation, confidence intervals for proportions and standard error.

Improved and streamlined discussion of finite population correction.

Updated Excel screenshots and functions.

Moved *MegaStat* and Minitab examples to end-of-chapter *Software Supplement*.

New, revised, and updated exercises.

Three new *LearningStats* demonstrations.

## Chapter 9—One-Sample Hypothesis Tests

Reorganized sections on Type I and Type II error, decision rules, and  $p$ -values).

Updated Excel screenshots and improved confidence interval figure.

Revised and updated examples (e.g., using software to reduce retail fraud).

Moved *MegaStat* and Minitab examples to end-of-chapter *Software Supplement*.

Excel functions for tests of proportions when normality cannot be assumed.

## Chapter 10—Two-Sample Hypothesis Tests

Simplified learning objectives to match content more closely.

Updated Excel screenshots and instructions.

Moved *MegaStat* and Minitab examples to end-of-chapter *Software Supplement*.

More emphasis on the question of whether or not sample sizes must be equal.

Improved notation for tests of two proportions and  $F$  tests.

New graphic for Excel  $F$  tests and corresponding Excel functions.

New, revised, and updated exercises (e.g., paired  $t$  tests).

## Chapter 11—Analysis of Variance

Improved notation and graphics to illustrate one-factor ANOVA (e.g., manufacturing defect rates).

Updated Excel screenshots and illustrations to emphasize Excel’s capabilities.

Moved *MegaStat* and Minitab examples to end-of-chapter *Software Supplement*.

Added an alternative formula for Hartley’s test.

Improved discussion of Tukey tests.

Optional section on Kruskal-Wallis test as alternative to ANOVA.

# THE THIRD EDITION

Clarified instructions on exercises to improve compatibility with Connect.

## Chapter 12—Simple Regression

Reorganized learning objectives to give more focus on testable topic.

Expanded discussion on the difference between association and cause and effect.

Updated Excel screenshots and illustrations to focus on Excel's capabilities.

Moved *MegaStat* and Minitab examples to end-of-chapter *Software Supplement*.

More explanation of interpreting the intercept.

Improved residual illustrations and new graphic on heteroscedasticity patterns.

Boxed comments on prediction interval width and unusual observations.

New, revised, and updated exercises (e.g., outliers, leverage, SAT scores, vehicle MPG, home values).

New section on logistic regression with a new logit data set and interpretive exercise.

Updated *Related Reading* and one new *Learning Stats* demonstration (correlation).

## Chapter 13—Multiple Regression

Greater emphasis on Excel, with most *MegaStat* and Minitab references moved to end-of-chapter *Software Supplement*.

Improved distinction between confidence intervals and prediction intervals.

Expanded discussion of multiplicative models and interaction effects.

New graphic on heteroscedasticity patterns.

Improved discussion of unusual observations.

New section on logistic regression with several predictors (complementing Chapter 12).

Expanded discussion of stepwise regression.

Revised exercise instructions for compatibility with Connect<sup>®</sup>.

Updated *Related Reading* and *LearningStats*.

## Chapter 14—Chi-Square Tests

Reorganized learning objectives to align with concepts and Connect<sup>®</sup>.

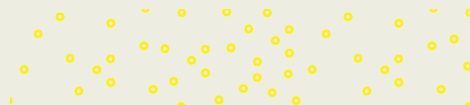
Updated screenshots, more Excel emphasis.

New graphics for GOF tests.

Streamlined discussion of topics (e.g., binning for normal GOF tests) and new graphics for ECDF tests.

New, revised, and updated data sets (e.g., Kentucky Derby, national league runs, U.S. presidents' ages).

Updated *Related Reading* references.



# BRIEF CONTENTS

- **CHAPTER ONE**  
Overview of Statistics 2
- **CHAPTER TWO**  
Data Collection 22
- **CHAPTER THREE**  
Describing Data Visually 56
- **CHAPTER FOUR**  
Descriptive Statistics 104
- **CHAPTER FIVE**  
Probability 166
- **CHAPTER SIX**  
Discrete Probability Distributions 206
- **CHAPTER SEVEN**  
Continuous Probability Distributions 242
- **CHAPTER EIGHT**  
Sampling Distributions and Estimation 280
- **CHAPTER NINE**  
One-Sample Hypothesis Tests 326
- **CHAPTER TEN**  
Two-Sample Hypothesis Tests 366
- **CHAPTER ELEVEN**  
Analysis of Variance 416
- **CHAPTER TWELVE**  
Simple Regression 444
- **CHAPTER THIRTEEN**  
Multiple Regression 508
- **CHAPTER FOURTEEN**  
Chi-Square Tests 564
- **APPENDIXES**
  - A Binomial Probabilities 598
  - B Poisson Probabilities 600
  - C-1 Standard Normal Areas 603
  - C-2 Cumulative Standard Normal Distribution 604
  - D Student's  $t$  Critical Values 606
  - E Chi-Square Critical Values 607
  - F Critical Values of  $F_{.10}$  608
  - G Solutions to Odd-Numbered Exercises 616
  - H Answers to Exam Review Questions 637
  - I Writing and Presenting Reports 639
  - J Excel Statistical Functions 644
- **INDEX 650**

# CONTENTS

## CHAPTER ONE

### Overview of Statistics 2

- 1.1 What Is Statistics? 3
- 1.2 Why Study Statistics? 5
- 1.3 Statistics in Business 7
- 1.4 Statistical Challenges 9
- 1.5 Critical Thinking 15
  - Chapter Summary 17
  - Chapter Exercises 18

## CHAPTER TWO

### Data Collection 22

- 2.1 Variables and Data 23
- 2.2 Level of Measurement 27
- 2.3 Sampling Concepts 31
- 2.4 Sampling Methods 34
- 2.5 Data Sources 43
- 2.6 Surveys 44
  - Chapter Summary 49
  - Chapter Exercises 50

## CHAPTER THREE

### Describing Data Visually 56

- 3.1 Stem-and-Leaf Displays and Dot Plots 57
- 3.2 Frequency Distributions and Histograms 62
- 3.3 Effective Excel Charts 70
- 3.4 Line Charts 71
- 3.5 Column and Bar Charts 75
- 3.6 Pie Charts 79
- 3.7 Scatter Plots 81
- 3.8 Tables 85
- 3.9 Deceptive Graphs 89
  - Chapter Summary 92
  - Chapter Exercises 93

## CHAPTER FOUR

### Descriptive Statistics 104

- 4.1 Numerical Description 105
- 4.2 Measures of Center 107
- 4.3 Measures of Variability 120
- 4.4 Standardized Data 128
- 4.5 Percentiles, Quartiles, and Box Plots 132
- 4.6 Covariance and Correlation 141
- 4.7 Grouped Data 146
- 4.8 Skewness And Kurtosis 148
  - Chapter Summary 152
  - Chapter Exercises 154

## CHAPTER FIVE

### Probability 166

- 5.1 Random Experiments 167
- 5.2 Probability 169
- 5.3 Rules of Probability 173
- 5.4 Independent Events 178
- 5.5 Contingency Tables 182
- 5.6 Tree Diagrams 189
- 5.7 Bayes' Theorem 191
- 5.8 Counting Rules 195
  - Chapter Summary 199
  - Chapter Exercises 200

## CHAPTER SIX

### Discrete Probability Distributions 206

- 6.1 Discrete Probability Distributions 207
- 6.2 Expected Value and Variance 210
- 6.3 Uniform Distribution 214
- 6.4 Binomial Distribution 216
- 6.5 Poisson Distribution 223
- 6.6 Hypergeometric Distribution 229
- 6.7 Transformations of Random Variables (Optional) 232
  - Chapter Summary 235
  - Chapter Exercises 236

## CHAPTER SEVEN

### Continuous Probability Distributions 242

- 7.1 Continuous Probability Distributions 243
- 7.2 Uniform Continuous Distribution 245
- 7.3 Normal Distribution 247
- 7.4 Standard Normal Distribution 250
- 7.5 Normal Approximations 263
- 7.6 Exponential Distribution 267
  - Chapter Summary 272
  - Chapter Exercises 274

## CHAPTER EIGHT

### Sampling Distributions and Estimation 280

- 8.1 Sampling and Estimation 281
- 8.2 Central Limit Theorem 285
- 8.3 Sample Size and Standard Error 290
- 8.4 Confidence Interval for a Mean ( $\mu$ ) with Known  $\sigma$  292
- 8.5 Confidence Interval for a Mean ( $\mu$ ) with Unknown  $\sigma$  295
- 8.6 Confidence Interval for a Proportion ( $\pi$ ) 302



## xxii Contents

- 8.7** Estimating From Finite Populations 308  
**8.8** Sample Size Determination for a Mean 310  
**8.9** Sample Size Determination for a Proportion 312  
**8.10** Confidence Interval for a Population Variance,  $\sigma^2$  (Optional) 314  
 Chapter Summary 316  
 Chapter Exercises 318
- 12.7** Confidence and Prediction Intervals For  $Y$  471  
**12.8** Residual Tests 474  
**12.9** Unusual Observations 480  
**12.10** Other Regression Topics (Optional) 483  
**12.11** Logistic Regression (Optional) 490  
 Chapter Summary 492  
 Chapter Exercises 494

**CHAPTER NINE****One-Sample Hypothesis Tests 326**

- 9.1** Logic of Hypothesis Testing 327  
**9.2** Type I and Type II Errors 330  
**9.3** Decision Rules and Critical Values 333  
**9.4** Testing a Mean: Known Population Variance 337  
**9.5** Testing a Mean: Unknown Population Variance 344  
**9.6** Testing a Proportion 350  
 Chapter Summary 359  
 Chapter Exercises 360

**CHAPTER TEN****Two-Sample Hypothesis Tests 366**

- 10.1** Two-Sample Tests 367  
**10.2** Comparing Two Means: Independent Samples 369  
**10.3** Confidence Interval for the Difference of Two Means,  $\mu_1 - \mu_2$  377  
**10.4** Comparing Two Means: Paired Samples 379  
**10.5** Comparing Two Proportions 386  
**10.6** Confidence Interval for the Difference of Two Proportions,  $\pi_1 - \pi_2$  393  
**10.7** Comparing Two Variances 394  
 Chapter Summary 401  
 Chapter Exercises 403

**CHAPTER ELEVEN****Analysis of Variance 416**

- 11.1** Overview of Anova 417  
**11.2** One-Factor Anova (Completely Randomized Model) 419  
**11.3** Multiple Comparisons 427  
**11.4** Tests for Homogeneity of Variances 429  
**11.5** Kruskal-Wallis Test (Optional) 433  
 Chapter Summary 434  
 Chapter Exercises 435

**CHAPTER TWELVE****Simple Regression 444**

- 12.1** Visual Displays and Correlation Analysis 445  
**12.2** Simple Regression 451  
**12.3** Regression Models 453  
**12.4** Ordinary Least Squares Formulas 457  
**12.5** Tests for Significance 461  
**12.6** Analysis of Variance: Overall Fit 467

**CHAPTER THIRTEEN****Multiple Regression 508**

- 13.1** Multiple Regression 509  
**13.2** Assessing Overall Fit 515  
**13.3** Predictor Significance 518  
**13.4** Confidence Intervals for  $Y$  522  
**13.5** Categorical Variables 525  
**13.6** Tests for Nonlinearity and Interaction 533  
**13.7** Multicollinearity 537  
**13.8** Regression Diagnostics 540  
**13.9** Other Regression Topics (Optional) 547  
 Chapter Summary 549  
 Chapter Exercises 551

**CHAPTER FOURTEEN****Chi-Square Tests 564**

- 14.1** Chi-Square Test for Independence 565  
**14.2** Chi-Square Tests for Goodness-of-Fit 576  
**14.3** Uniform Goodness-of-Fit Test 579  
**14.4** Normal Chi-Square Goodness-of-Fit Test 583  
**14.5** ECDF Tests (Optional) 586  
 Chapter Summary 587  
 Chapter Exercises 588

**APPENDIXES**

- A** Binomial Probabilities 598  
**B** Poisson Probabilities 600  
**C-1** Standard Normal Areas 603  
**C-2** Cumulative Standard Normal Distribution 604  
**D** Student's  $t$  Critical Values 606  
**E** Chi-Square Critical Values 607  
**F** Critical Values of  $F_{.10}$  608  
**G** Solutions to Odd-Numbered Exercises 616  
**H** Answers to Exam Review Questions 637  
**I** Writing and Presenting Reports 639  
**J** Excel Statistical Functions 644

**INDEX 650**



# Essential Statistics

**in Business and Economics**

*Third Edition*

## CHAPTER

## 2

## Data Collection

## CHAPTER CONTENTS

- 2.1 Variables and Data
- 2.2 Level of Measurement
- 2.3 Sampling Concepts
- 2.4 Sampling Methods
- 2.5 Data Sources
- 2.6 Surveys

## CHAPTER LEARNING OBJECTIVES

## LO

When you finish this chapter you should be able to

- LO 2-1** Use basic terminology for describing data and samples.
- LO 2-2** Explain the difference between numerical and categorical data.
- LO 2-3** Explain the difference between time series and cross-sectional data.
- LO 2-4** Recognize levels of measurement in data and ways of coding data.
- LO 2-5** Recognize a Likert scale and know how to use it.
- LO 2-6** Use the correct terminology for samples and populations.
- LO 2-7** Explain the common sampling methods and how to implement them.
- LO 2-8** Find everyday print or electronic data sources.
- LO 2-9** Describe basic elements of survey types, survey designs, and response scales.





©DreamPictures/Blend Images LLC

In scientific research, data arise from experiments whose results are recorded systematically. In business, data usually arise from accounting transactions or management processes (e.g., inventory, sales, payroll). Much of the data that statisticians analyze were observed and recorded without explicit consideration of their statistical uses, yet important decisions may depend on the data. How many pints of type A blood will be required at Mt. Sinai Hospital next Thursday? How many dollars must State Farm keep in its cash account to cover automotive accident claims next November? How many yellow three-quarter-sleeve women's sweaters will Lands' End sell this month? To answer such questions, we usually look at historical, or empirical, data.<sup>1</sup>

## 2.1 VARIABLES AND DATA

### Data Terminology

An **observation** is a single member of a collection of items that we want to study, such as a person, firm, or region. An example of an observation is an employee or an invoice mailed last month. A **variable** is a characteristic of the subject or individual, such as an employee's income or an invoice amount. The **data set** consists of all the values of all of the variables for all of the observations we have chosen to observe. In this book, we will use **data** as a plural, and data set to refer to a collection of observations taken as a whole. Data usually are entered into a spreadsheet or database as an  $n \times m$  matrix. Specifically, each column is a variable ( $m$  columns), and each row is an observation ( $n$  rows). Table 2.1 shows a small data set with eight observations (8 rows) and five variables (5 columns).


A data set may consist of many variables. The questions that can be explored and the analytical techniques that can be used will depend upon the data type and the number of variables. This textbook starts with **univariate data sets** (one variable), then moves to **bivariate data sets** (two variables) and **multivariate data sets** (more than two variables), as illustrated in Table 2.2.

#### LO 2-1

Use basic terminology for describing data and samples.

<sup>1</sup>Is *data* singular or plural? *Data* is the plural of the Latin *datum* (a "given" fact). But in the popular press you will often see "data" used synonymously with "information" and hence as a singular ("The compressed data is stored on a CD . . .").

**TABLE 2.1**

**A Small Multivariate Data Set (5 variables, 8 observations)**  
 **SmallData**

Obs	Name	Age	Income	Position	Gender	Education
1	Frieda	45	\$ 67,100	Personnel director	F	Master's
2	Stefan	32	56,500	Operations analyst	M	Doctorate
3	Barbara	55	88,200	Marketing VP	F	Master's
4	Donna	27	59,000	Statistician	F	Bachelor's
5	Larry	46	36,000	Security guard	M	High School
6	Alicia	52	68,500	Comptroller	F	Master's
7	Alec	65	95,200	Chief executive	M	Bachelor's
8	Jaime	50	71,200	Public relations	M	Bachelor's

**TABLE 2.2**

**Number of Variables and Typical Tasks**

Data Set	Variables	Example	Typical Tasks
Univariate	One	Income	Histograms, basic statistics
Bivariate	Two	Income, Age	Scatter plots, correlation
Multivariate	More than two	Income, Age, Gender	Regression modeling

**LO 2-2**

Explain the difference between numerical and categorical data.

**Categorical and Numerical Data**

A data set may contain a mixture of *data types*. Two broad categories are *categorical data* and *numerical data*, as shown in Figure 2.1.

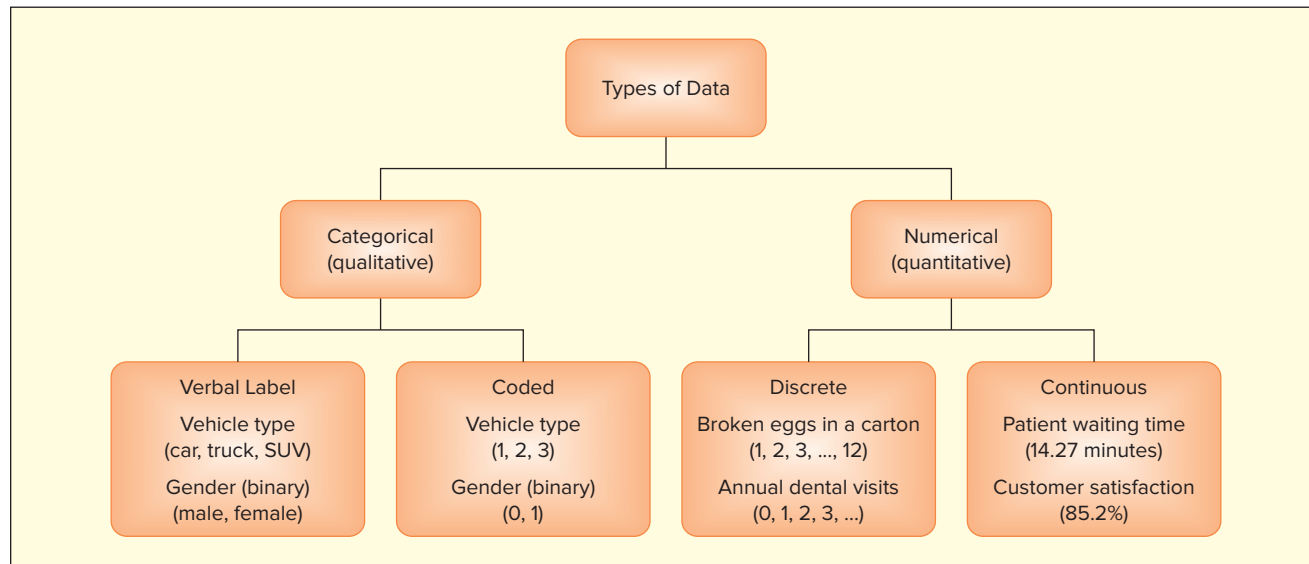
**Categorical data** (also called *qualitative data*) have values that are described by words rather than numbers. For example, structural lumber can be classified by the lumber type (e.g., fir, hemlock, pine), automobile styles can be classified by size (e.g., full, midsize, compact, subcompact), and movies can be categorized using common movie classifications (e.g., action and adventure, children and family, classics, comedy, documentary).

Because categorical variables have nonnumerical values, it might seem that categorical data would be of limited statistical use. In fact, there are many statistical methods that can handle categorical data, which we will introduce in later chapters. On occasion, the values of the categorical variable might be represented using numbers. This is called **coding**. For example, a database might code payment methods using numbers:

1 = cash    2 = check    3 = credit/debit card    4 = gift card

**FIGURE 2.1**

**Data Types and Examples**



Coding a category as a number does *not* make the data numerical and the numbers do not typically imply a rank. But on occasion, a ranking does exist. For example, a database might code education degrees using numbers:

1 = Bachelor's    2 = Master's    3 = Doctorate

Some categorical variables have only two values. We call these **binary variables**. Examples include employment status (e.g., employed or unemployed), mutual fund type (e.g., load or no-load), and marital status (e.g., currently married or not currently married). Binary variables are often coded using a 1 or 0. For a binary variable, the 0-1 coding is arbitrary, so the choice is equivalent. For example, a variable such as gender could be coded as:

1 = female    0 = male

or as

1 = male    0 = female

**Numerical data** (also called *quantitative* data) arise from counting, measuring something, or some kind of mathematical operation. For example, we could count the number of auto insurance claims filed in March (e.g., 114 claims) or sales for last quarter (e.g., \$4,920), or we could measure the amount of snowfall over the last 24 hours (e.g., 3.4 inches). Most accounting data, economic indicators, and financial ratios are quantitative, as are physical measurements.

Numerical data can be broken down into two types. A variable with a countable number of distinct values is **discrete**. Often, such data are integers. You can recognize integer data because their description begins with “number of.” For example, the number of Medicaid patients in a hospital waiting room (e.g., 2) or the number of takeoffs at Chicago O’Hare International Airport in an hour (e.g., 37). Such data are integer variables because we cannot observe a fractional number of patients or takeoffs.

A numerical variable that can have any value within an interval is **continuous**. This would include things like physical measurements (e.g., distance, weight, time, speed) or financial variables (e.g., sales, assets, price/earnings ratios, inventory turns); for example, runner Usain Bolt’s time in the 100-meter dash (e.g., 9.58 seconds) or the weight of a package of Sun-Maid raisins (e.g., 427.31 grams). These are continuous variables because any interval such as [425, 429] grams can contain infinitely many possible values. Sometimes we round a continuous measurement to an integer (e.g., 427 grams), but that does not make the data discrete.

Ambiguity between discrete and continuous is introduced when we round continuous data to whole numbers (e.g., your weight this morning). Consider a package of Sun-Maid raisins that is labeled 425 grams. The underlying measurement scale is continuous (e.g., on an accurate scale, its weight might be 425.31 grams, a noninteger). Precision depends on the instrument we use to measure the continuous variable. Shoes are sized in discrete half steps (e.g., 6, 6.5, 7) even though the length of your foot is a continuous variable. Conversely, we sometimes treat discrete data as continuous when the range is very large (e.g., SAT scores) and when small differences (e.g., 604 or 605) aren’t of much importance. We generally treat financial data (dollars, euros, pesos) as continuous even though retail prices go in discrete steps of .01 (i.e., we go from \$1.25 to \$1.26). This topic will be discussed in later chapters. If in doubt, just think about how  $X$  was measured and whether or not its values are countable.

## Time Series Data and Cross-Sectional Data

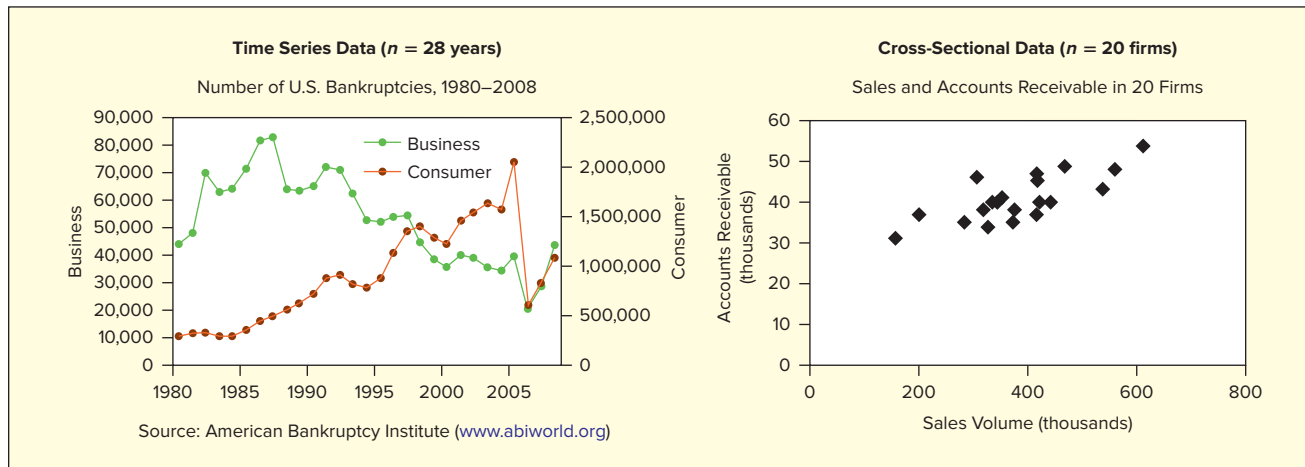
If each observation in the sample represents a different equally spaced point in time (years, months, days), we have **time series data**. The *periodicity* is the time between observations. It may be annual, quarterly, monthly, weekly, daily, hourly, etc. Examples of *macroeconomic* time series data would include national income (GDP, consumption, investment), economic indicators (Consumer Price Index, unemployment rate, Standard & Poor’s 500 Index), and monetary data (M1, M2, M3, prime rate, T-bill rate, consumer borrowing, federal debt). Examples of *microeconomic* time series data would include a firm’s sales, market share, debt/equity ratio, employee absenteeism, inventory turnover, and product quality ratings. For time series, we are interested in *trends and patterns over time* (e.g., personal bankruptcies from 1980 to 2008 as shown in Figure 2.2).

### LO 2-3

Explain the difference between time series and cross-sectional data.

FIGURE 2.2

## Examples of Time Series versus Cross-Sectional Data



Source: American Bankruptcy Institute ([www.abiworld.org](http://www.abiworld.org)).

If each observation represents a different individual unit (e.g., a person, firm, geographic area) at the same point in time, we have **cross-sectional data**. Thus, traffic fatalities in the 50 U.S. states for a given year, debt/equity ratios for the *Fortune* 500 firms in the last quarter of a certain year, last month's Visa balances for a bank's new mortgage applicants, or GPAs of students in a statistics class would be cross-sectional data. For cross-sectional data, we are interested in *variation among observations* (e.g., accounts receivable in 20 Subway franchises) or in *relationships* (e.g., whether accounts receivable are related to sales volume in 20 Subway franchises as shown in Figure 2.2).

Some variables (such as unemployment rates) could be either time series (monthly data over each of 60 months) or cross-sectional (January's unemployment rate in 50 different cities). We can combine the two (e.g., monthly unemployment rates for the 13 Canadian provinces or territories for the last 60 months) to obtain *pooled cross-sectional and time series data*.

## SECTION EXERCISES



- 2.1 What type of data (categorical, discrete numerical, or continuous numerical) is each of the following variables? If there is any ambiguity about the data type, explain why the answer is unclear.
  - a. The manufacturer of your car.
  - b. Your college major.
  - c. The number of college credits you are taking.
- 2.2 What type of data (categorical, discrete numerical, or continuous numerical) is each of the following variables? If there is any ambiguity, explain why the answer is unclear.
  - a. Length of a TV commercial.
  - b. Number of peanuts in a can of Planters Mixed Nuts.
  - c. Occupation of a mortgage applicant.
  - d. Flight time from London Heathrow to Chicago O'Hare.
- 2.3 What type of data (categorical, discrete numerical, or continuous numerical) is each of the following variables? If there is any ambiguity about the data type, explain why the answer is unclear.
  - a. The miles on your car's odometer.
  - b. The fat grams you ate for lunch yesterday.
  - c. The name of the airline with the cheapest fare from New York to London.
  - d. The brand of cell phone you own.
- 2.4 (a) Give three original examples of discrete data. (b) Give three original examples of continuous data. In each case, explain and identify any ambiguities that might exist. *Hint:* Do not restrict yourself to published data. Consider data describing your own life (e.g., your sports performance, financial data, or academic data). You need *not* list all the data, merely describe them and show a few typical data values.

- 2.5 Which type of data (cross-sectional or time series) is each variable?
- Scores of 50 students on a midterm accounting exam last semester.
  - Bob's scores on 10 weekly accounting quizzes last semester.
  - Average score by all takers of the state's CPA exam for each of the last 10 years.
  - Number of years of accounting work experience for each of the 15 partners in a CPA firm.
- 2.6 Which type of data (cross-sectional or time series) is each variable?
- Value of Standard & Poor's 500 stock price index at the close of each trading day last year.
  - Closing price of each of the 500 stocks in the S&P 500 index on the last trading day this week.
  - Dividends per share paid by General Electric common stock for the last 20 quarters.
  - Latest price/earnings ratios of 10 stocks in Bob's retirement portfolio.
- 2.7 Which type of data (cross-sectional or time series) is each variable?
- Mexico's GDP for each of the last 10 quarters.
  - Unemployment rates in each of the 31 states in Mexico at the end of last year.
  - Unemployment rate in Mexico at the end of each of the last 10 years.
  - Average home value in each of the 10 largest Mexican cities today.
- 2.8 Give an original example of a time series variable and a cross-sectional variable. Use your own experience (e.g., your sports activities, finances, education).

## 2.2 LEVEL OF MEASUREMENT

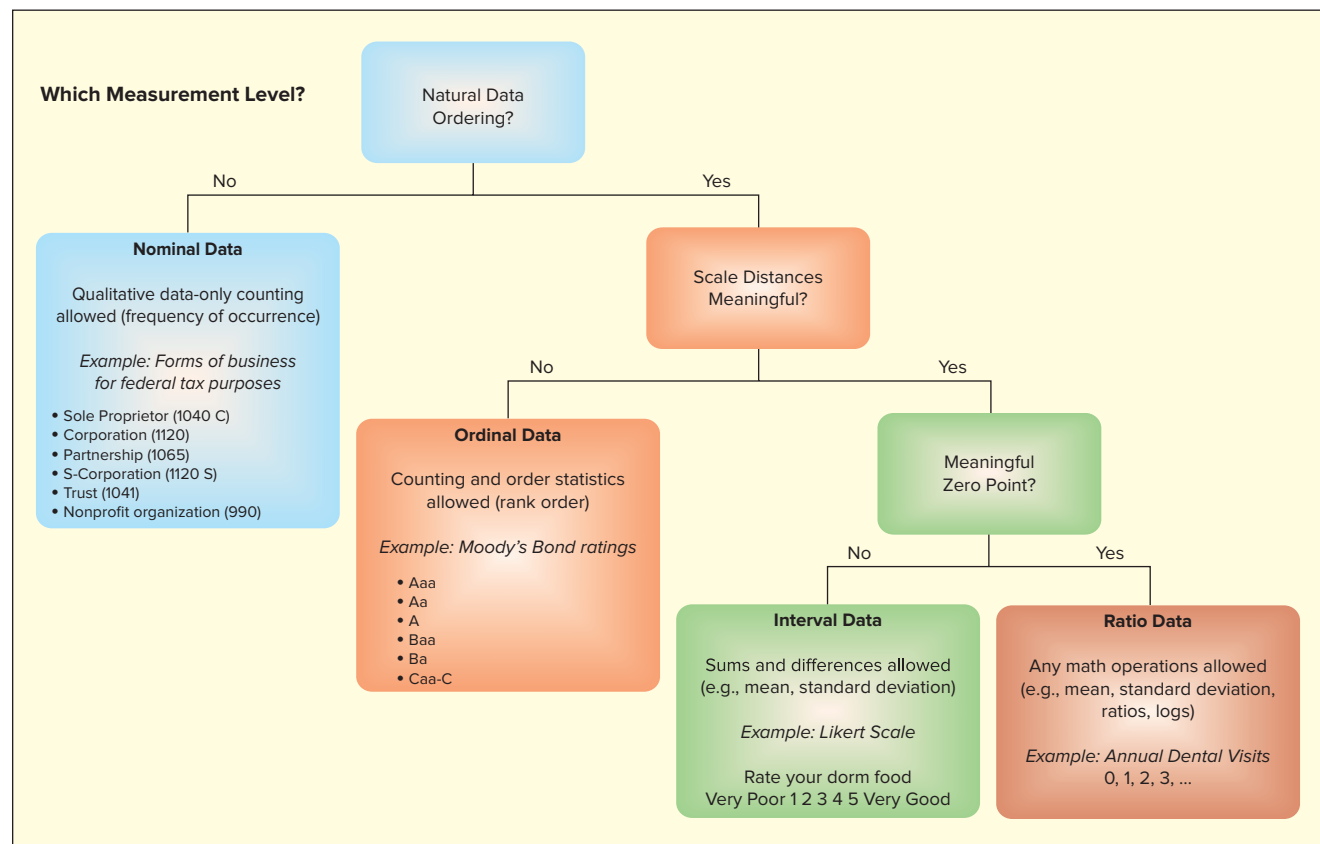
Data types shown in Figure 2.1 can be further classified by their measurement level. Statisticians typically refer to four levels of measurement for data: nominal, ordinal, interval, and ratio. This typology was proposed over 60 years ago by psychologist S. S. Stevens. The allowable mathematical operations, statistical summary measures, and statistical tests depend on the measurement level. The criteria are summarized in Figure 2.3.

### LO 2-4

Recognize levels of measurement in data and ways of coding data.

FIGURE 2.3

### Determining the Measurement Level





## Nominal Measurement

Nominal measurement is the weakest level of measurement and the easiest to recognize. **Nominal data** (from Latin *nomen*, meaning “name”) merely identify a *category*. “Nominal” data are the same as “qualitative,” “categorical,” or “classification” data. To be sure that the categories are collectively exhaustive, it is common to use **Other** as the last item on the list. For example, the following survey questions yield nominal data:

Did you file an insurance claim last month?

1. Yes
2. No

Which cell phone service provider do you use?

1. AT&T
2. Sprint-Nextel
3. T-Mobile
4. Verizon
5. Other

We usually code nominal data numerically. However, the codes are arbitrary placeholders with no numerical meaning, so it is improper to perform mathematical analysis on them. For example, we would not calculate an average using the cell phone service data (1 through 5). This may seem obvious, yet people have been known to do it. Once the data are in the computer, it’s easy to forget that the “numbers” are only categories. With nominal data, the only permissible mathematical operations are counting (e.g., frequencies) and a few simple statistics such as the mode.

## Ordinal Measurement

**Ordinal data** codes connote a *ranking* of data values. For example:

What size automobile do you usually drive?

1. Full-size
2. Compact
3. Subcompact

How often do you use Microsoft Access?

1. Frequently
2. Sometimes
3. Rarely
4. Never

Thus, a 2 (Compact) implies a larger cars than a 3 (Subcompact). Like nominal data, these ordinal numerical codes lack the properties that are required to compute many statistics, such as the average. Specifically, there is no clear meaning to the *distance* between 1 and 2, or between 2 and 3, or between 3 and 4 (What would be the distance between “Rarely” and “Never”?). Other examples of ordinal scales can be found in a recruiter’s rating of job candidates (outstanding, good, adequate, weak, unsatisfactory), S&P credit ratings (AAA, AA+, AA, AA–, A+, A, A–, B+, B, B–, etc.) or job titles (president, group vice president, plant manager, department head, clerk). Ordinal data can be treated as nominal, but not vice versa. Ordinal data are especially common in social sciences, marketing, and human resources research. There are many useful statistical tests for ordinal data.

### LO 2-5

Recognize a Likert scale and know how to use it.

## Interval Measurement

The next step up the measurement scale is **interval data**. Interval data are used frequently and are important in business. Interval data often arise from surveys where customers are asked to rate their satisfaction with a service or product on a numerical scale. While these scale points are expressed as numbers (e.g., 1–10), the scale is arbitrary, the numbers are not a count or a physical measure, and the value “0” has no meaning. However, we can say that the distances between scale points have meaning. The difference between a rating of 4 and 6 is treated the same as the difference between 7 and 9. Other examples include the Celsius or Fahrenheit scales of temperature. Because intervals between numbers represent distances, we can do mathematical operations such as taking an average. But because the zero point of these scales is arbitrary, we can’t say that a customer who rates our service an 8 is twice as satisfied as a customer who rates our service a 4. Nor can we say that 60°F is twice as warm as 30°F. That is, ratios are not meaningful for interval data. The absence of a meaningful zero is a key characteristic of interval data.

The **Likert scale** is a special case that is frequently used in survey research. You have undoubtedly seen such scales. Typically, a statement is made, and the respondent is asked to indicate his or her agreement/disagreement on a five-point or seven-point scale using verbal anchors. The *coarseness* of a Likert scale refers to the number of scale points (typically 5 or 7). For example:

“College-bound high school students should be required to study a foreign language.” (check one)

Strongly Somewhat Neither Agree Somewhat Strongly  
Agree Agree nor Disagree Disagree Disagree

A neutral midpoint (“Neither Agree nor Disagree”) is allowed if we use an *odd* number of scale points (usually 5 or 7). Occasionally, surveys may omit the neutral midpoint to force the respondent to “lean” one way or the other. Likert data are coded numerically (e.g., 1 to 5), but any equally spaced values will work, as shown in Table 2.3.

Likert Coding: 1 to 5 Scale	Likert Coding: -2 to +2 Scale
5 = Will help a lot	+2 = Will help a lot
4 = Will help a little	+1 = Will help a little
3 = No effect on investment climate	0 = No effect on investment climate
2 = Will hurt a little	-1 = Will hurt a little
1 = Will hurt a lot	-2 = Will hurt a lot

**TABLE 2.3**

**Examples of Likert-Scale Coding: “How will deflation affect the investment climate?”**

But do Likert data qualify as interval measurements? By choosing the verbal anchors carefully, many researchers believe that the *intervals* are the same (e.g., the distance from 1 to 2 is “the same” as the *interval*, say, from 3 to 4). However, ratios are not meaningful (i.e., here 4 is not twice 2). The assumption that Likert scales produce interval data justifies a wide range of statistical calculations, including averages, correlations, and so on. Researchers use many Likert-scale variants.

How would you rate your Internet service provider? (check one.)

Terrible  Poor  Adequate  Good  Excellent

Instead of labeling every response category, marketing surveys might put verbal anchors only on the end points. This avoids intermediate scale labels and permits any number of scale points. Likert data usually are discrete, but some web surveys now use a continuous response scale that allows the respondent to position a “slider” anywhere along the scale to produce continuous data (actually, the number of positions is finite but very large). For example:

Likert (using discrete scale points)                      Likert (using a slider)

Very Poor 1 2 3 4 5 6 7 Very Good                      Very Poor  Very Good

### Ratio Measurement

Ratio measurement is the strongest level of measurement. **Ratio data** have all the properties of the other three data types, but in addition possess a *meaningful zero* that represents the absence of the quantity being measured. Because of the zero point, ratios of data values are meaningful (e.g., \$20 million in profit is twice as much as \$10 million). Balance sheet data, income statement data, financial ratios, physical counts, scientific measurements, and most engineering measurements are ratio data because zero has meaning (e.g., a company with zero sales sold nothing). Having a zero point does *not* restrict us to positive data. For example, profit is a ratio variable (e.g., \$4 million is twice \$2 million), yet firms can have negative profit (i.e., a loss).

Zero does *not* have to be observable in the data. Newborn babies, for example, cannot have zero weight, yet baby weight clearly is ratio data (i.e., an 8-pound baby is 33 percent heavier than a 6-pound baby). What matters is that the zero is an absolute reference point. The Kelvin

temperature scale is a ratio measurement because its absolute zero represents the absence of molecular vibration, while zero on the Celsius scale is merely a convenience (note that  $30^{\circ}\text{C}$  is not “twice as much temperature” as  $15^{\circ}\text{C}$ ).

Lack of a true zero is often the quickest test to defrock variables masquerading as ratio data. For example, a Likert scale (+2, +1, 0, -1, -2) is *not* ratio data despite the presence of zero because the zero (neutral) point does not connote the absence of anything. As an acid test, ask yourself whether 2 (strongly agree) is twice as much “agreement” as 1 (slightly agree). Some classifications are debatable. For example, college GPA has a zero, but does it represent the absence of learning? Does 4.00 represent “twice as much” learning as 2.00? Is there an underlying reality ranging from 0 to 4 that we are measuring? Most people seem to think so, although the conservative procedure would be to limit ourselves to statistical tests that assume only ordinal data.

Although beginning statistics textbooks usually emphasize interval or ratio data, there are textbooks that emphasize other kinds of data, notably in behavioral research (e.g., psychology, sociology, marketing, human resources).

We can recode ratio measurements *downward* into ordinal or nominal measurements (but not conversely). For example, doctors may classify systolic blood pressure as “normal” (under 130), “elevated” (130 to 140), or “high” (140 or over). The recoded data are ordinal because the ranking is preserved. Intervals may be unequal. For example, U.S. air traffic controllers classify planes as “small” (under 41,000 pounds), “large” (41,001 to 254,999 pounds), and “heavy” (255,000 pounds or more). Such recoding is done to simplify the data when the exact data magnitude is of little interest; however, we discard information if we map stronger measurements into weaker ones.

## SECTION EXERCISES

 connect

- 2.9** Which measurement level (nominal, ordinal, interval, ratio) is each of the following variables? Explain.
- Number of hits in Game 1 of the next World Series.
  - Baltimore’s relative standing in the American League East (among five teams).
  - Field position of a randomly chosen baseball player (catcher, pitcher, etc.).
  - Temperature on opening day (Celsius).
  - Salary of a randomly chosen American League pitcher.
  - Freeway traffic on opening day (light, medium, heavy).
- 2.10** Which measurement level (nominal, ordinal, interval, ratio) is each of the following variables? Explain.
- Number of employees in the Walmart store in Hutchinson, Kansas.
  - Number of merchandise returns on a randomly chosen Monday at a Walmart store.
  - Temperature (in Fahrenheit) in the ice-cream freezer at a Walmart store.
  - Name of the cashier at register 3 in a Walmart store.
  - Manager’s rating of the cashier at register 3 in a Walmart store.
  - Social Security number of the cashier at register 3 in a Walmart store.
- 2.11** Which measurement level (nominal, ordinal, interval, ratio) is each of the following variables? Explain.
- Number of passengers on Delta Flight 833.
  - Waiting time (minutes) after gate pushback before Delta Flight 833 takes off.
  - Brand of cell phone owned by a cabin attendant on Delta Flight 833.
  - Ticket class (first, business, or economy) of a randomly chosen passenger on Delta Flight 833.
  - Outside air temperature (Celsius) when Delta Flight 833 reaches 35,000 feet.
  - Passenger rating (on five-point Likert scale) of Delta’s in-flight food choices.
- 2.12** Which measurement level (nominal, ordinal, interval, ratio) is the response to each question? If you think that the level of measurement is ambiguous, explain why.
- How would you describe your level of skill in using Excel? (check one)
    - Low    Medium    High
  - How often do you use Excel? (check one)
    - Rarely    Often    Very Often
  - Which version of Excel for Windows do you use? (check one)
    - 2007    2010    2013    2016
  - I spend \_\_\_\_\_ hours a day using Excel.



- 2.13** Here is a question from a ski resort guest satisfaction survey that uses a five-point scale. (a) Would the measurement level for the data collected from this question be nominal, ordinal, interval, or ratio? (b) Would it be appropriate to calculate an average rating for the various items? Explain. (c) Would a 10-point scale be better? Explain.

“Rate your satisfaction level on numerous aspects of *today’s* experience, where 1 = Extremely Dissatisfied and 5 = Extremely Satisfied.”

1. Value for Price Paid:	1	2	3	4	5
2. Ticket Office Line Wait (if went to ticket window):	1	2	3	4	5
3. Friendliness/Helpfulness of Lift Operators:	1	2	3	4	5
4. Lift Line Waits:	1	2	3	4	5
5. Ski Patrol Visibility:	1	2	3	4	5

- 2.14** (a) Would the measurement level for the data collected from this Microsoft® survey question be nominal, ordinal, interval, or ratio? (b) Would a “6” response be considered twice as good as a “3” response? Why or why not? (c) Would a 1–5 scale be adequate? Explain.

**Microsoft® Quality of Support Survey**

Please rate the overall quality of support you received from Microsoft on this particular issue, using a 9-point scale where 9 is Excellent and 1 is Very Poor.

Excellent
Very Poor
Don't Know

9 8 7 6 5 4 3 2 1

Source: Microsoft Corporation

## 2.3 SAMPLING CONCEPTS

There are almost 2 million retail businesses in the United States. It is unrealistic for market researchers to study them all or in a timely way. But since 2001, a new firm called ShopperTrak RCT ([www.shoppertrak.com](http://www.shoppertrak.com)) has been measuring purchases at a sample of 45,000 mall-based stores and using this information to advise clients quickly of changes in shopping trends. This application of sampling is part of the relatively new field of *retail intelligence*. In this section, you will learn the differences between a **sample** and a **population**, and why sometimes a sample is necessary or desirable.

### LO 2-6

Use the correct terminology for samples and populations.

### Population or Sample?

<b>Population</b>	All of the items that we are interested in. May be either finite (e.g., all of the passengers on a particular plane) or effectively infinite (e.g., all of the Cokes produced in an ongoing bottling process).
<b>Sample</b>	A subset of the population that we will actually analyze.

### Sample or Census?

A *sample* involves looking only at some items selected from the population, while a **census** is an examination of all items in a defined population. The accuracy of a census can be illusory. For example, the 2000 U.S. decennial census is believed to have overcounted by 1.3 million people while the 2010 census count is thought to have overestimated the U.S. population by only 36,000. Reasons include the extreme mobility of the U.S. population and the fact that

some people do not want to be found (e.g., illegal immigrants) or do not reply to the mailed census form. Further, budget constraints make it difficult to train enough census field workers, install data safeguards, and track down incomplete responses or nonresponses. For these reasons, U.S. censuses have long used sampling in certain situations. Many statistical experts advised using sampling more extensively in the 2000 decennial census, but the U.S. Congress concluded that an actual headcount must be attempted.

When the quantity being measured is volatile, there cannot be a census. For example, The Arbitron Company tracks American radio listening habits using over 2.6 million “Radio Diary Packages.” For each “listening occasion,” participants note start and stop times for each station. Panelists also report their age, sex, and other demographic information. Table 2.4 outlines some situations where a sample rather than a census would be preferred, and vice versa.

TABLE 2.4

## Sample or Census?

<i>Situations Where a Sample May Be Preferred</i>	<i>Situations Where a Census May Be Preferred</i>
<p><b>Infinite Population</b></p> <p>No census is possible if the population is of indefinite size (an assembly line can keep producing bolts, a doctor can keep seeing more patients).</p>	<p><b>Small Population</b></p> <p>If the population is small, there is little reason to sample, for the effort of data collection may be only a small part of the total cost.</p>
<p><b>Destructive Testing</b></p> <p>The act of measurement may destroy or devalue the item (battery life, vehicle crash tests).</p>	<p><b>Large Sample Size</b></p> <p>If the required sample size approaches the population size, we might as well go ahead and take a census.</p>
<p><b>Timely Results</b></p> <p>Sampling may yield more timely results (checking wheat samples for moisture content, checking peanut butter for salmonella contamination).</p>	<p><b>Database Exists</b></p> <p>If the data are on disk, we can examine 100% of the cases. But auditing or validating data against physical records may raise the cost.</p>
<p><b>Accuracy</b></p> <p>Instead of spreading resources thinly to attempt a census, our budget might be better spent to improve training of field interviewers and improve data safeguards.</p>	<p><b>Legal Requirements</b></p> <p>Banks must count <i>all</i> the cash in bank teller drawers at the end of each business day. The U.S. Congress forbade sampling in the 2000 decennial population census.</p>
<p><b>Cost</b></p> <p>Even if a census is feasible, the cost, in either time or money, may exceed our budget.</p>	
<p><b>Sensitive Information</b></p> <p>A trained interviewer might learn more about sexual harassment in an organization through confidential interviews of a small sample of employees.</p>	

## Parameters and Statistics

From a sample of  $n$  items, chosen from a population, we compute **statistics** that can be used as estimates of **parameters** found in the population. To avoid confusion, we use different symbols for each parameter and its corresponding statistic. Thus, the population mean is denoted  $\mu$  (the lowercase Greek letter mu) while the sample mean is  $\bar{x}$ . The population proportion is denoted  $\pi$  (the lowercase Greek letter pi), while the sample proportion is  $p$ . Figure 2.4 illustrates this idea.

For example, suppose we want to know the mean (average) repair cost for auto air-conditioning warranty claims or the proportion (percent) of 25-year-old concertgoers who have permanent hearing loss. Because a census is impossible, these parameters would be estimated using a sample. For the sample statistics to provide good estimates of the population parameters, the population

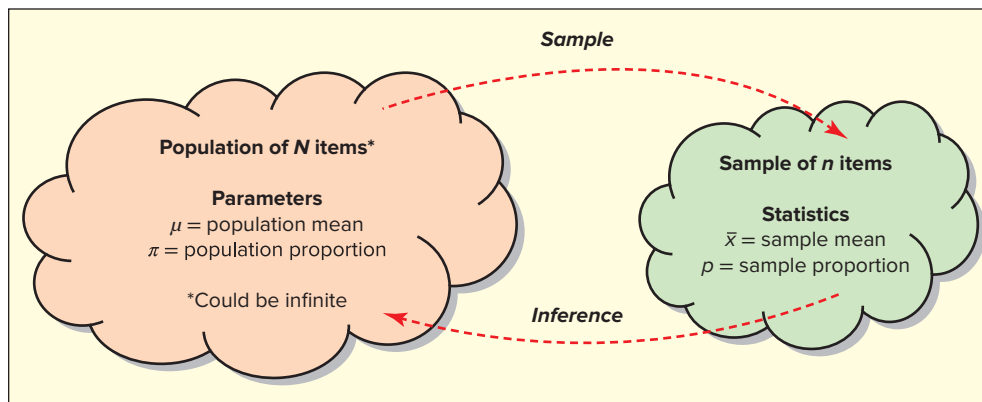


FIGURE 2.4

Population versus Sample

## Parameter or Statistic?

<b>Parameter</b>	A measurement or characteristic of the population (e.g., a mean or proportion). Usually unknown because we can rarely observe the entire population. Usually (but not always) represented by a Greek letter (e.g., $\mu$ or $\pi$ ).
<b>Statistic</b>	A numerical value calculated from a sample (e.g., a mean or proportion). Usually (but not always) represented by a Roman letter (e.g., $\bar{x}$ or $p$ ).

must be carefully specified, and the sample must be drawn scientifically so the sample items are representative of the population.

## Target Population

A population may be defined either by a list (e.g., the names of the passengers on Flight 234) or by a rule (e.g., the customers who eat at Noodles & Company). The **target population** contains all the individuals in which we are interested. Suppose we wish to estimate the proportion of potential consumers who would purchase a \$20 Harley-Davidson desk calendar. Is the target population all drivers? Only male drivers over age 16? Only drivers with incomes over \$25,000? Only motorcycle owners? By answering questions such as these, we not only identify the target population but also are forced to define our business goals more clearly. The **sampling frame** is the group from which we take the sample. If the frame differs from the target population, then our estimates might not be accurate. Examples of frames are phone directories, voter registration lists, alumni association mailing lists, or marketing databases. Other examples might be:

- Names and addresses of all registered voters in Colorado Springs, Colorado.
- Names and addresses of all vehicle owners in Ventura County, California.
- E-mail addresses of all L.L.Bean customers who have placed online orders.

The sample for the U.S. Energy Information Administration's survey of gasoline prices is drawn from a frame of approximately 115,000 retail gasoline outlets, constructed from purchased private commercial sources and EIA sources, combined with zip codes from private lists. Individual frames are mapped to the county level by using zip codes, and outlets are assigned to standard metropolitan statistical areas from Census Bureau definitions. (For details, see [www.eia.doe.gov](http://www.eia.doe.gov).)

## EXAMPLE 2.1

*Gasoline Price Survey*

## Mini Case

2.1

## How Many People Watch the Super Bowl Each Year?

The Super Bowl, the annual championship football game between the NFC and AFC football conferences, is the most-watched television program every year. The game is broadcast on national TV, and viewership has been steadily growing since the game was first played on January 15, 1967. Nielsen Media Research collects viewing information for the Super Bowl each year. In 2016 (Denver Broncos versus Carolina Panthers), Nielsen reported the Super Bowl had a 46.6 rating and an average 111.9 million viewers. These numbers were lower than those for 2015 (New England Patriots versus Seattle Seahawks), which saw a 49.7 rating and an average 114.4 million viewers. The rating estimates the percentage of households with televisions that are watching the Super Bowl and in 2015/2016 Nielsen reported that 116,400,000 U.S. households had televisions. The average number of viewers is an estimate of the average number of people watching at any given minute during the broadcast.

Both of these measures are estimates based on samples. How are these samples collected? Nielsen collects data from a sample of households using both a viewing diary, where viewers manually track the programs they've watched, and a meter that is connected to the TV set. The sample sizes are in the magnitude of 25,000 households and 50,000 people. The people who make up these samples come from different age groups, income levels, and geography. The quality of the sampling process is important because much of these data are used to guide decisions about advertising. And by some estimates, advertising is a \$70 billion industry! The Nielsen sampling process is overseen by the Media Ratings Council and is audited each year by the public accounting firm Ernst & Young. More information on their sampling techniques can be found at [www.nielsen.com](http://www.nielsen.com).

## SECTION EXERCISES



- 2.15 Would you use a sample or a census to measure each of the following? Why?
- The model years of the cars driven by each of your five closest friends.
  - The model years of the cars driven by each student in your statistics class.
  - The model years of the cars driven by each student in your university.
  - The model years of the cars driven by each professor whose classes you are taking.
- 2.16 Would you use a sample or a census to measure each of the following? Why? If you are uncertain, explain the issues.
- The mean time battery life of your laptop computer in continuous use.
  - The number of students in your statistics class who brought laptop computers to class today.
  - The average price paid for a laptop computer by students at your university.
  - The percentage of disk space available on laptop computers owned by your five closest friends.
- 2.17 The target population is all stocks in the S&P 500 index. Is each of the following a parameter or a statistic?
- The average price/earnings ratio for all 500 stocks in the S&P index.
  - The proportion of all stocks in the S&P 500 index that had negative earnings last year.
  - The proportion of energy-related stocks in a random sample of 50 stocks.
  - The average rate of return for 20 stocks recommended by a broker.

## LO 2-7

Explain the common sampling methods and how to implement them.

## 2.4 SAMPLING METHODS

There are two main categories of sampling methods. In **random sampling**, items are chosen by randomization or a chance procedure. The idea of random sampling is to produce a sample that is representative of the population. **Nonrandom sampling** is less scientific but is sometimes used for expediency.

## Random Sampling Methods

We will first discuss the four random sampling techniques shown in Table 2.5 and then describe three commonly used nonrandom sampling techniques, summarized in Table 2.8.

Simple random sample	Use random numbers to select items from a list (e.g., Visa cardholders).
Systematic sample	Select every $k$ th item from a list or sequence (e.g., restaurant customers).
Stratified sample	Select randomly within defined strata (e.g., by age, occupation, gender).
Cluster sample	Select random geographical regions (e.g., zip codes) that represent the population.

**TABLE 2.5**
**Random Sampling Methods**

We denote the population size by  $N$  and the sample size by  $n$ . In a **simple random sample**, every item in the population of  $N$  items has the same chance of being chosen in the sample of  $n$  items. A physical experiment to accomplish this would be to write each of the  $N$  data values on a poker chip, and then to draw  $n$  chips from a bowl after stirring it thoroughly. But we can accomplish the same thing if the  $N$  population items appear on a numbered list, by choosing  $n$  integers between 1 and  $N$  that we match up against the numbered list of the population items.

For example, suppose we want to select one student at random from a list of 15 students (see Figure 2.5). If you were asked to “use your judgment,” you would probably pick a name in the middle, thereby biasing the draw against those individuals at either end of the list. Instead we rely on a **random number** to “pick” the name. How do we determine the random number? Before computers, statisticians relied on published tables of random numbers. The process is simpler today. Even most pocket calculators have a key to produce a random decimal in the interval  $[0, 1]$  that can be converted to a random integer. In this example, we used Excel’s function =RANDBETWEEN(1,15) to pick a random integer between 1 and 15. The number was 12, so Stephanie was selected. There is no bias because all values from 1 to 15 are *equiprobable* (i.e., equally likely to occur).

Random person <b>12</b>					
1	Adam	6	Haitham	11	Moira
2	Addie	7	Jackie	<b>12</b>	<b>Stephanie</b>
3	Don	8	Judy	13	Stephen
4	Floyd	9	Lindsay	14	Tara
5	Gadis	10	Majda	15	Xander

**FIGURE 2.5**
**Picking on Stephanie**

**Sampling without replacement** means that once an item has been selected to be included in the sample, it cannot be considered for the sample again. The Excel function =RANDBETWEEN(a,b) uses **sampling with replacement**. This means that the same random number could show up more than once. Using the bowl analogy, if we throw each chip back in the bowl and stir the contents before the next draw, an item can be chosen again. Instinctively, most people believe that sampling without replacement is preferred over sampling with replacement because allowing duplicates in our sample seems odd. In reality, sampling without replacement can be a problem when our sample size  $n$  is close to our population size  $N$ . At some point in the sampling process, the remaining items in the population will no longer have the same probability of being selected as the items we chose at the beginning of the sampling process. This could lead to a bias (a tendency to overestimate or underestimate the parameter we are trying to measure) in our sample results. Sampling with replacement does not lead to bias.

When should we worry about sampling without replacement? Only when the population is finite and the sample size is close to the population size. Consider the Russell 3000® Index, which

has 3,000 stocks. If you sample 100 stocks, without replacement, you have “used” only about 3 percent of the population. The sample size  $n = 100$  is considered small relative to the population size  $N = 3,000$ . A common criterion is that a finite population is *effectively infinite* if the sample is less than 5 percent of the population (i.e., if  $n/N \leq .05$ ). In Chapter 8, you will learn how to adjust for the effect of population size when you make a sample estimate. For now, you only need to recognize that such adjustments are of little consequence when the population is large.

## Infinite Population?

When the sample is less than 5 percent of the population (i.e., when  $n/N \leq .05$ ), then the population is effectively infinite. An equivalent statement is that a population is effectively infinite when it is at least 20 times as large as the sample (i.e., when  $N/n \geq 20$ ).

Because computers are easier, we rarely use random number tables. Table 2.6 shows a few alternative ways to choose 10 integers between 1 and 875. All are based on a software algorithm that creates uniform decimal numbers between 0 and 1. Excel’s function `=RAND()` does this, and many pocket calculators have a similar function. We call these *pseudorandom* generators because even the best algorithms eventually repeat themselves (after a cycle of millions of numbers). Thus, a software-based random data encryption scheme could conceivably be broken. To enhance data security, Intel and other firms are examining hardware-based methods (e.g., based on thermal noise or radioactive decay) to prevent patterns or repetition. Fortunately, most applications don’t require that degree of randomness. For example, choosing shuffle for your playlist does not generate strictly random songs because its random numbers are generated by an algorithm from a “seed number” that eventually repeats. However, the repeat period is so great that an iPod user would never notice. Excel’s and Minitab’s random numbers are good enough for most purposes.

**TABLE 2.6**

**Some Ways to Get 10 Random Integers between 1 and 875**

Excel—Option A	Enter the Excel function <code>=RANDBETWEEN(1,875)</code> into 10 spreadsheet cells. Press F9 to get a new sample.
Excel—Option B	Enter the function <code>=INT(1+875*RAND())</code> into 10 spreadsheet cells. Press F9 to get a new sample.
Internet	The website <a href="http://www.random.org">www.random.org</a> will give you many kinds of excellent random numbers (integers, decimals, etc).
Minitab	Use Minitab’s Random Data menu with the Integer option.
Pocket Calculator	Press the RAND key to get a random number in the interval $[0, 1]$ , multiply by 875, then round up to the next integer.

## Randomizing a List

To randomize a list (assuming it is in a spreadsheet), we can insert the Excel function `=RAND()` beside each row. This creates a column of random decimal numbers between 0 and 1. Copy the random numbers and paste them in the same column using Paste Special > Values to “fix” them (otherwise they will keep changing). Then sort all the columns by the random number column, and *voilà*—the list is now random! The first  $n$  items on the randomized list can now be used as a random sample. This method is especially useful when the list is very long (perhaps millions of lines). The first  $n$  items are a random sample of the entire list, for they are as likely as any others.

Another method of random sampling is to choose every  $k$ th item from a sequence or list, starting from a randomly chosen entry among the first  $k$  items on the list. This is called **systematic sampling**. Figure 2.6 shows how to sample every fourth item, starting from item 2, resulting in a sample of  $n = 20$  items.



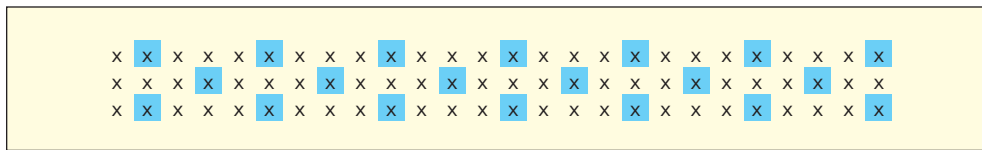


FIGURE 2.6

Systematic Sampling

An attraction of systematic sampling is that it can be used with unlistable or infinite populations, such as production processes (e.g., testing every 5,000th light bulb) or political polling (e.g., surveying every 10th voter who emerges from the polling place). Systematic sampling is also well-suited to linearly organized physical populations (e.g., pulling every 10th patient folder from alphabetized filing drawers in a veterinary clinic).

A systematic sample of  $n$  items from a population of  $N$  items requires that periodicity  $k$  be approximately  $N/n$ . For example, to choose 25 companies from a list of 500 companies in Example 2.2 (Table 2.7), we chose every twentieth stock ( $k = 500/25 = 20$ ).

To sample the compensation of the CEOs of the 500 largest companies in the United States listed in *Forbes'* annual survey, take every 20th company in the alphabetized list, starting (randomly) with the 13th company. The starting point (the 13th company) is chosen at random. This yields the sample of 25 CEOs, shown in Table 2.7. While it would be very time-consuming to examine all 500 executives, this sample should provide a representative cross-section.

## EXAMPLE 2.2

CEO Compensation

Observation	Firm	CEO	One-Year Total (\$mil)
1	AK Steel Holding	James L. Wainscott	11.82
2	Anadarko Petroleum	James T. Hackett	19.65
3	Avnet	Roy Vallee	10.16
4	Bristol-Myers Squibb	James M. Cornelius	5.06
5	Charter Commun	Neil Smit	5.63
6	Commercial Metals	Murray R. McClean	3.84
7	CVS Caremark	Thomas M. Ryan	19.55
8	Dynegy	Bruce A. Williamson	8.70
9	Estee Lauder Cos	William P. Lauder	5.32
10	FPL Group	Lewis Hay III	14.25
11	Google	Eric E. Schmidt	0.48
12	Huntington Bancshs	Thomas E. Hoaglin	0.98
13	Johnson Controls	Stephen A. Roell	15.69
14	Leucadia National	Ian M. Cumming	1.21
15	MBIA	Joseph W. Brown	22.20
16	Morgan Stanley	John J. Mack	17.65
17	Northeast Utilities	Charles W. Shivery	5.91
18	People's United	Philip R. Sherringham	2.22
19	Progress Energy	William D. Johnson	4.11
20	Rockwell Collins	Clayton M. Jones	11.31
21	Sovereign Bancorp	Joseph P. Campanelli	2.48
22	TD Ameritrade Holding	Joseph H. Moglia	3.76
23	Union Pacific	James R. Young	7.19
24	Wal-Mart Stores	H Lee Scott Jr	8.65
25	Wynn Resorts	Stephen A. Wynn	11.25

TABLE 2.7

## CEO Compensation in 25 Large U.S. Firms

Source: Forbes.com, April 30, 2008.  
Compensation is for 2007.

Systematic sampling should yield acceptable results unless patterns in the population happen to recur at periodicity  $k$ . For example, weekly pay cycles ( $k = 7$ ) would make it illogical to sample bank check cashing volume every Friday. A less obvious example would be a machine that stamps a defective part every 12th cycle due to a bad tooth in a 12-tooth gear, which would make it misleading to rely on a sample of every 12th part ( $k = 12$ ). But periodicity coincident with  $k$  is not typical or expected in most situations.

Sometimes we can improve our sample efficiency by utilizing prior information about the population. This method is applicable when the population can be divided into relatively homogeneous subgroups of known size (called *strata*). Within each *stratum*, a simple random sample of the desired size could be taken. Alternatively, a random sample of the whole population could be taken, and then individual strata estimates could be combined using appropriate weights. This procedure, called **stratified sampling**, can reduce cost per observation and narrow the error bounds. For a population with  $L$  strata, the population size  $N$  is the sum of the stratum sizes:  $N = N_1 + N_2 + \dots + N_L$ . The weight assigned to stratum  $j$  is  $w_j = N_j/N$  (i.e., each stratum is weighted by its known proportion of the population).

To illustrate, suppose we want to estimate MMR (measles-mumps-rubella) vaccination rates among employees in state government, and we know that our target population (those individuals we are trying to study) is 55 percent male and 45 percent female. Suppose our budget only allows a sample of size 200. To ensure the correct gender balance, we could sample 110 males and 90 females. Alternatively, we could just take a random sample of 200 employees. Although our random sample probably will not contain *exactly* 110 males and 90 females, we can get an overall estimate of vaccination rates by *weighting* the male and female sample vaccination rates using  $w_M = 0.55$  and  $w_F = 0.45$  to reflect the known strata sizes.

## Mini Case

## 2.2

### Sampling for Safety

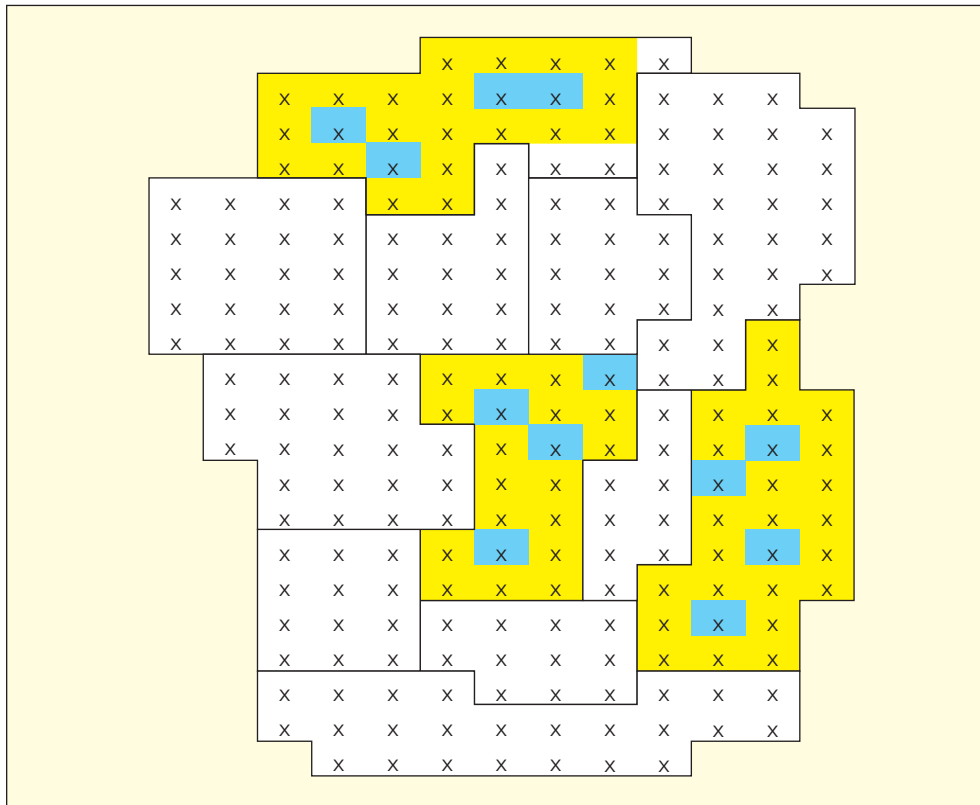
To help automakers and other researchers study the causes of injuries and fatalities in vehicle accidents, the U.S. Department of Transportation developed the National Accident Sampling System (NASS) Crashworthiness Data System (CDS). Because it is impractical to investigate every accident (there were 6,159,000 police-reported accidents in 2005), detailed data are collected in a common format from 24 primary sampling units, chosen to represent all serious police-reported motor vehicle accidents in the United States during the year. Selection of sample accidents is done in three stages: (1) The country is divided into 1,195 geographic areas called Primary Sampling Units (PSUs) grouped into 12 strata based on geographic region. Two PSUs are selected from each stratum using weights roughly proportional to the number of accidents in each stratum. (2) In each sampled PSU, a second stage of sampling is performed using a sample of Police Jurisdictions (PJs) based on the number, severity, and type of accidents in the PJ. (3) The final stage of sampling is the selection of accidents within the sampled PJs. Each reported accident is classified into a stratum based on type of vehicle, most severe injury, disposition of the injured, tow status of the vehicles, and model year of the vehicles. Each team is assigned a fixed number of accidents to investigate each week, governed by the number of researchers on a team. Weights for the strata are assigned to favor a larger percentage of higher severity accidents while ensuring that accidents in the same stratum have the same probability of being selected, regardless of the PSU. The NASS CDS database is administered by the National Center for Statistics and Analysis (NCSA) of the National Highway Traffic Safety Administration (NHTSA). These data are currently helping to improve the government's "5 Star" crashworthiness rating system for vehicles.

Source: See: [www.nrd.nhtsa.dot.gov/Pubs/NASS94.PDF](http://www.nrd.nhtsa.dot.gov/Pubs/NASS94.PDF)

**Cluster samples** are taken from strata consisting of geographical regions. We divide a region (say, a city) into subregions (say, blocks, subdivisions, or school districts). In one-stage cluster sampling, our sample consists of all elements in each of  $k$  randomly chosen subregions



(or clusters). In two-stage cluster sampling, we first randomly select  $k$  subregions (clusters) and then choose a random sample of elements within each cluster. Figure 2.7 illustrates how four elements could be sampled from each of three randomly chosen clusters using two-stage cluster sampling.

**FIGURE 2.7**

**Two-Stage Cluster Sampling: Randomly choose three clusters, then randomly choose four items in each cluster**

Because elements within a cluster are proximate, travel time and interviewer expenses are kept low. Cluster sampling is useful when:

- Population frame and stratum characteristics are not readily available.
- It is too expensive to obtain a simple or stratified sample.
- The cost of obtaining data increases sharply with distance.
- Some loss of reliability is acceptable.

Although cluster sampling is cheap and quick, it is often reasonably accurate because people in the same neighborhood tend to be similar in income, ethnicity, educational background, and so on. Cluster sampling is useful in political polling, surveys of gasoline pump prices, studies of crime victimization, vaccination surveys, or lead contamination in soil. A hospital may contain clusters (floors) of similar patients. A warehouse may have clusters (pallets) of inventory parts. Forest sections may be viewed as clusters to be sampled for disease or timber growth rates.

Cluster sampling is also widely used in marketing and economic surveys. The Bureau of Labor Statistics relies on multistage cluster sampling for estimating economic indicators such as the Consumer Price Index (CPI) and employment rates. The CPI measures the average change in price for a “market basket” of goods and services typically used by urban consumers. The CPI is estimated from a two-stage cluster sampling process. The sampling process begins with 87 urban areas in the United States. Within these urban areas, prices on over 200 categories are gathered from approximately 50,000 housing units and 23,000 retail establishments.

## Nonrandom Sampling Methods

Table 2.8 describes three commonly used nonrandom sampling techniques. Businesses often rely on these techniques to quickly gather data that might be used to guide informal decisions or as preliminary data to help design formal studies that use random samples.

**TABLE 2.8**

**Nonrandom Samples**

Judgment sample	Use expert knowledge to choose “typical” items (e.g., which employees to interview).
Convenience sample	Use a sample that happens to be available (e.g., ask co-workers’ opinions at lunch).
Focus groups	In-depth dialog with a representative panel of individuals (e.g., iPod users).

**Judgment sampling** is a nonrandom sampling method that relies on the expertise of the sampler to choose items that are representative of the population. For example, to estimate the corporate spending on research and development (R&D) in the medical equipment industry, we might ask an industry expert to select several “typical” firms. Unfortunately, subconscious biases can affect experts, too. In this context, “bias” does not mean prejudice, but rather *non-randomness* in the choice. Judgment samples may be the best alternative in some cases, but we can’t be sure whether the sample was random. *Quota sampling* is a special kind of judgment sampling in which the interviewer chooses a certain number of people in each category (e.g., men/women).

The sole virtue of **convenience sampling** is that it is quick. The idea is to grab whatever sample is handy. An accounting professor who wants to know how many MBA students would take a summer elective in international accounting can just survey the class she is currently teaching. The students polled may not be representative of all MBA students, but an answer (although imperfect) will be available immediately. A newspaper reporter doing a story on perceived airport security might interview co-workers who travel frequently. An executive might ask department heads if they think nonbusiness web surfing is widespread.

You might think that convenience sampling is rarely used or, when it is, that the results are used with caution. However, this does not appear to be the case. Because convenience samples often sound the first alarm on a timely issue, their results have a way of attracting attention and have probably influenced quite a few business decisions. The mathematical properties of convenience samples are unknowable, but they do serve a purpose and their influence cannot be ignored.

A **focus group** is a panel of individuals chosen to be representative of a wider population, formed for open-ended discussion and idea gathering about an issue (e.g., a proposed new product or marketing strategy). Typically 5–10 people are selected, and the interactive discussion lasts one to two hours. Participants are usually individuals who do not know each other, but who are prescreened to be broadly compatible yet diverse. A trained moderator guides the focus group’s discussion and keeps it on track. Although not a random sampling method, focus groups are widely used, both in business and in social science research, for the insights they can yield beyond “just numbers.”

## Other Data Collection Methods

Businesses now have many other methods for collecting data with the use of technology. Point-of-sale (POS) systems can collect real-time data on purchases at retail or convenience stores, restaurants, and gas stations. Many companies use loyalty cards that are swiped during the purchase. These loyalty cards have the customer’s information, which can be matched to the purchase just made. Businesses also send out e-mail surveys to loyal customers on a regular basis to get feedback on their products and services. Facebook can track your Internet searches using its software algorithms. Google also tracks Internet searches and provides these data through its Google Analytics services.

## Mini Case

## 2.3

### Pricing Accuracy

Bar code price scanning using the Universal Product Code (UPC) became the standard in most retail businesses following the rapid improvement in scanning technology during the 1970s. Since that time, federal and state agencies have monitored businesses to regulate pricing accuracy at their checkouts. Because a census is impossible for checking price accuracy, sampling is an essential tool in enforcing consumer protection laws. The National Institute for Standards and Technology (NIST) has developed a handbook for inspection agencies that provides guidance on how to conduct a pricing sampling inspection.

Arizona's Department of Weights and Measures (DWM) has set up a UPC scanner pricing sampling inspection process for the retail establishments in that state. A UPC inspection will be based on either a stratified sample (e.g., a cosmetics department) or a simple random sample from throughout the store. The inspector will select between 25 and 50 items based on the sample size recommendation from NIST. The items will be taken to the register for scanning and the inspector will count the number of items that show a difference between the display and scanned price. Arizona requires that the retail store have 98 percent accuracy.

Between 2001 and 2006, in the state of Arizona, Walmart failed 526 price accuracy inspections. The Arizona attorney general filed a lawsuit against Walmart in 2006. The lawsuit was settled when Walmart agreed to a financial settlement of \$1 million and modifications of its pricing practices.

Source: See <https://www.azag.gov/press-release/terry-goddard-announces-1-million-pricing-settlement-wal-mart>.

## Sample Size

The necessary sample size depends on the inherent variability of the quantity being measured and the desired precision of the estimate. For example, the caffeine content of Mountain Dew is fairly consistent because each can or bottle is filled at the factory, so a small sample size would suffice to estimate the mean. In contrast, the amount of caffeine in an individually brewed cup of Bigelow Raspberry Royale tea varies widely because people let it steep for varying lengths of time, so a larger sample would be needed to estimate the mean. The purposes of the investigation, the costs of sampling, the budget, and time constraints also are taken into account in deciding on sample size. Setting the sample size is worth a detailed discussion, found in later chapters.

## Sources of Error or Bias

In sampling, the word *bias* does not refer to prejudice. Rather, it refers to a systematic tendency to over- or underestimate a population parameter of interest. However, the words *bias* and *error* are often used interchangeably. The word *error* generally refers to problems in sample methodology that lead to inaccurate estimates of a population parameter. No matter how careful you are when conducting a survey, you will encounter potential sources of error. Let's briefly review a few, summarized in Table 2.9.

Source of Error	Characteristics
Nonresponse bias	Respondents differ from nonrespondents
Selection bias	Self-selected respondents are atypical
Response error	Respondents give false information
Coverage error	Incorrect specification of frame or population
Measurement error	Unclear survey instrument wording
Interviewer error	Responses influenced by interviewer
Sampling error	Random and unavoidable

**TABLE 2.9**

**Potential Sources of Survey Error**

**Nonresponse bias** occurs when those who respond have characteristics different from those who don't respond. For example, people with caller ID, answering machines, blocked or unlisted numbers, or cell phones are likely to be missed in telephone surveys. Because these are generally more affluent individuals, their socioeconomic class may be underrepresented in the poll. A special case is **selection bias**, a self-selected sample. For example, a talk show host who invites viewers to take a web survey about their sex lives will attract plenty of respondents. But those who are willing to reveal details of their personal lives (and who have time to complete the survey) are likely to differ substantially from those who dislike nosy surveys or are too busy (and probably weren't watching the show anyway).


Further, it is easy to imagine that hoax replies will be common to such a survey (e.g., a bunch of college dorm students giving silly answers on a web survey). **Response error** occurs when respondents deliberately give false information to mimic socially acceptable answers, to avoid embarrassment, or to protect personal information.

**Coverage error** occurs when some important segment of the target population is systematically missed. For example, a survey of Notre Dame University alumni will fail to represent noncollege graduates or those who attended public universities. **Measurement error** results when the survey questions do not accurately reveal the construct being assessed. When the interviewer's facial expressions, tone of voice, or appearance influences the responses, data are subject to **interviewer error**.

Finally, **sampling error** is uncontrollable random error that is inherent in any random sample. Even when using a random sampling method, it is possible that the sample will contain unusual responses. This cannot be prevented and is generally undetectable. It is *not* an error on your part.

## SECTION EXERCISES

 connect

- 2.18 The target population is all students in your university. You wish to estimate the average current Visa balance for each student. How large would the university student population have to be in order to be regarded as effectively infinite in each of the following samples?
- A sample of 10 students.
  - A sample of 50 students.
  - A sample of 100 students.
- 2.19 Suppose you want to know the ages of moviegoers who attend the latest *X-Men* movie. What kind of sample is it if you (a) survey the first 20 persons to emerge from the theater, (b) survey every 10th person to emerge from the theater, and (c) survey everyone who looks under age 12?
- 2.20 Suppose you want to study the number of e-mail accounts owned by students in your statistics class. What kind of sample is it if you (a) survey each student who has a student ID number ending in an odd number, (b) survey all the students sitting in the front row, and (c) survey every fifth student who arrives at the classroom?
- 2.21 Below is a  $6 \times 8$  array containing the ages of moviegoers (see file  **X-Men**). Treat this as a population. Select a random sample of 10 moviegoers' ages by using (a) simple random sampling with a random number table, (b) simple random sampling with Excel's =RANDBETWEEN() function, (c) systematic sampling, (d) judgment sampling, and (e) convenience sampling. Explain your methods.

32	34	33	12	57	13	58	16
23	23	62	65	35	15	17	20
14	11	51	33	31	13	11	58
23	10	63	34	12	15	62	13
40	11	18	62	64	30	42	20
21	56	11	51	38	49	15	21

- 2.22 (a) In the previous population, what was the proportion of all 48 moviegoers who were under age 30? (b) For each of the samples of size  $n = 10$  that you took, what was the proportion of moviegoers under age 30? (c) Was each sample proportion close to the population proportion?
- 2.23 In Excel, type a list containing names for 10 of your friends into cells B1:B10. Choose three names at random by randomizing this list. To do this, enter =RAND() into cells A1:A10, copy the random column and paste it using Paste Special > Values to fix the random numbers, and then sort the list by the random column. The first three names are the random sample.

## 2.5 DATA SOURCES

One goal of a statistics course is to help you learn where to find data that might be needed. Fortunately, many excellent sources are widely available, either in libraries or through private purchase. Table 2.10 summarizes a few of them.

### LO 2-8

Find everyday print or electronic data sources.

Type of Data	Examples
U.S. job-related data	U.S. Bureau of Labor Statistics
U.S. economic data	<i>Economic Report of the President</i>
Almanacs	<i>World Almanac, Time Almanac</i>
Periodicals	<i>Economist, Bloomberg Businessweek, Fortune, Forbes</i>
Indexes	<i>The New York Times, The Wall Street Journal</i>
Databases	Compustat, Citibase, U.S. Census
World data	<i>CIA World Factbook</i>
Web	Google, Yahoo!, MSN

**TABLE 2.10**

**Useful Data Sources**

The U.S. Census Bureau and the U.S. Bureau of Labor Statistics are rich sources of data on many different aspects of life in the United States. The publications library supported by the Census Bureau can be found at [www.census.gov](http://www.census.gov). The monthly, quarterly, and annual reports published by the Bureau of Labor statistics can be found at [www.bls.gov](http://www.bls.gov). It should be noted that until 2012, the *Statistical Abstract of the United States* was the largest, most general freely available annual compendium of facts and figures from public sources. A 2012 review of publications sponsored by the U.S. Census Bureau concluded with a decision to quit publishing the *Statistical Abstract*. However, you can still access previous years' publications at the U.S. Census website.

For annual and monthly time series economic data, try the *Economic Report of the President (ERP)*, which is published every February. The tables in the *ERP* can be downloaded for free in Excel format. Data on cities, counties, and states can be found in the *State and Metropolitan Area Data Book*, published every few years by the Bureau of the Census and available on CD-ROM in many libraries.

Annual almanacs from several major publishers are sold at most bookstores. These include data reprinted from the above sources, but also information on recent events, sports, the stock market, elections, Congress, world nations, states, and higher education. One of these almanacs should be on every informed citizen's shelf.

Annual surveys of major companies, markets, and topics of business or personal finance are found in magazines such as *Bloomberg Businessweek*, *Consumer Reports*, *Forbes*, *Fortune*, and *Money*. Indexes such as the *Business Periodical Index*, *The New York Times Index*, and *The Wall Street Journal Index* are useful for locating topics. Libraries have web search engines that can access many of these periodicals in abstract or full-text form.

Specialized computer databases (e.g., CRSP, Compustat, Citibase, U.S. Census) are available (at a price) for research on stocks, companies, financial statistics, and census data. An excellent summary of sources is F. Patrick Butler's *Business Research Sources: A Reference Navigator*. The web allows us to use search engines (e.g., Google, Yahoo!, MSN) to find information. Sometimes you may get lucky, but web information is often undocumented, unreliable, or unverifiable. Better information is available through private companies or trade associations, though often at a steep price. Related Reading can be found at the end of this chapter and Web Data Sources are listed below.

Often overlooked sources of help are your university librarians. University librarians understand how to find databases and how to navigate databases quickly and accurately. Librarians can help you distinguish between valid and invalid internet sources and then help you put the source citation in the proper format when writing reports.

## Web Data Sources

<i>Source</i>	<i>Website</i>
Bureau of Economic Analysis	<a href="http://www.bea.gov">www.bea.gov</a>
Bureau of Justice Statistics	<a href="http://www.bjs.gov">www.bjs.gov</a>
Bureau of Labor Statistics	<a href="http://www.bls.gov">www.bls.gov</a>
Central Intelligence Agency	<a href="http://www.cia.gov">www.cia.gov</a>
Economic Report of the President	<a href="http://www.gpo.gov/erp">www.gpo.gov/erp</a>
Environmental Protection Agency	<a href="http://www.epa.gov">www.epa.gov</a>
Federal Reserve System	<a href="http://www.federalreserve.gov">www.federalreserve.gov</a>
Food and Drug Administration	<a href="http://www.fda.gov">www.fda.gov</a>
National Agricultural Statistics Service	<a href="http://www.nass.usda.gov">www.nass.usda.gov</a>
National Center for Education Statistics	<a href="http://nces.ed.gov">nces.ed.gov</a>
National Center for Health Statistics	<a href="http://www.cdc.gov/nchs">www.cdc.gov/nchs</a>
State and Metropolitan Area Data Book	<a href="http://www.census.gov/library/publications/2010/compendia/databooks.html">www.census.gov/library/publications/2010/compendia/databooks.html</a>
Statistics Canada	<a href="http://www.statcan.gc.ca">www.statcan.gc.ca</a>
U.N. Department of Economic and Social Affairs	<a href="http://www.un.org/depts/unsd">www.un.org/depts/unsd</a>
U.S. Census Bureau	<a href="http://www.census.gov">www.census.gov</a>
U.S. Federal Statistics	<a href="http://fedstats.sites.usa.gov">fedstats.sites.usa.gov</a>
World Bank	<a href="http://www.worldbank.org">www.worldbank.org</a>
World Demographics	<a href="http://www.demographia.com">www.demographia.com</a>
World Health Organization	<a href="http://www.who.int/en">www.who.int/en</a>

### LO 2-9

Describe basic elements of survey types, survey designs, and response scales.

## 2.6 SURVEYS

Most survey research follows the same basic steps. These steps may overlap in time:

- Step 1: State the goals of the research.
- Step 2: Develop the budget (time, money, staff).
- Step 3: Create a research design (target population, frame, sample size).
- Step 4: Choose a survey type and method of administration.
- Step 5: Design a data collection instrument (questionnaire).
- Step 6: Pretest the survey instrument and revise as needed.
- Step 7: Administer the survey (follow up if needed).
- Step 8: Code the data and analyze it.

### Survey Types

Surveys fall into five general categories: mail, telephone, interview, web, and direct observation. They differ in cost, response rate, data quality, time required, and survey staff training requirements. Table 2.11 lists some common types of surveys and a few of their salient strengths/weaknesses.

**Response Rates** Consider the *cost per valid response*. A telephone survey might be cheapest to conduct, but bear in mind that over half the households in some metropolitan areas have unlisted phones, and many have answering machines or call screening. The sample you get may not be very useful in terms of reaching the target population. Telephone surveys (even with random dialing) do lend themselves nicely to cluster sampling (e.g., using each three-digit area code as a cluster and each three-digit exchange as a cluster) to sample somewhat homogeneous populations. Similarly, mail surveys can be clustered by zip code, which is a significant attraction. Web surveys are cheap, but rather uncontrolled. Nonresponse bias is a problem with all of these. Interviews or observational experiments are expensive and labor-intensive, but they



Survey Type	Characteristics
Mail	Mail requires a well-targeted and current mailing list (people move a lot). Expect low response rates and nonresponse bias (nonrespondents differ from those who respond). Zip code lists (often costly) are an attractive option to define strata of similar income, education, and attitudes. To encourage participation, a cover letter should explain the uses of the survey data. Plan for follow-up mailings.
Telephone	Random dialing yields low response and is poorly targeted. Purchased phone lists help reach the target population, though a low response rate still is typical (disconnected phones, caller screening, answering machines, work hours, no-call lists). Other sources of nonresponse bias include the growing number of cell phones, non-English speakers, and distrust caused by scams and robocalls.
Interviews	Interviewing is expensive and time-consuming, yet a trade-off between sample size for high-quality results may be worth it. Interviewers must be well-trained—an added cost. Interviewers can obtain information on complex or sensitive topics (e.g., gender discrimination in companies, birth control practices, diet and exercise).
Web	Web surveys are growing in popularity but are subject to nonresponse bias because they miss those who feel too busy, don't own computers, or distrust your motives (scams and spam). This type of survey works best when targeted to a well-defined interest group on a question of self-interest (e.g., views of CPAs on Sarbanes-Oxley accounting rules, frequent flyer views on airline security).

TABLE 2.11

## Common Types of Surveys

may provide higher quality data. Large-scale national research projects (e.g., mental health status of U.S. household members) offer financial incentives to encourage participants who otherwise would not provide information. Research suggests that adjustments can be made for whatever biases may result from such incentives. Table 2.12 offers some tips to conduct successful surveys.

Planning	What is the survey's purpose? What do you really need to know? What staff expertise is available? What skills are best hired externally? What degree of precision is required? What is your budget?
Design	To ensure a good response and useful data, you must invest time and money in designing the survey. Take advantage of many useful books and references so that you do not make unnecessary errors.
Quality	Care in preparation is needed. Glossy printing and advertising have raised people's expectations about quality. A scruffy questionnaire will be ignored. Some surveys (e.g., web-based) may require special software.
Pilot Test	Questions that are clear to you may be unclear to others. You can pretest the questionnaire on friends or co-workers, but using a small test panel of naive respondents who don't owe you anything is best.
Buy-In	Response rates may be improved by stating the purpose of the survey, by offering a token of appreciation (e.g., discount coupon, free gift), or with endorsements (e.g., from a trusted professional group).
Expertise	Consider hiring a consultant at the early stages, even if you plan to do your own data collection and tabulation. Early consultation is more effective than waiting until you get in trouble.

TABLE 2.12

## Survey Guidelines


## Questionnaire Design

You should consider hiring a consultant, at least in the early stages, to help you get your survey off the ground successfully. Alternatively, resources are available on the web to help you plan a survey. The American Statistical Association ([www.amstat.org](http://www.amstat.org)) offers brochures *What Is a Survey* and *How to Plan a Survey*. Additional materials are available from the Research Industry Coalition, Inc. ([www.researchindustry.org](http://www.researchindustry.org)), and the Insights Association ([www.insightsassociation.org](http://www.insightsassociation.org)). Entire books have been written to help you design and administer your own survey (see Related Reading).

The layout must not be crowded (use lots of white space). Begin with very short, clear instructions, stating the purpose, assuring anonymity, and explaining how to submit the completed survey. Questions should be numbered. Divide the survey into sections if the topics fall naturally into distinct areas. Let respondents bypass sections that aren't relevant to them (e.g., "If you answered no to Question 7, skip directly to Question 15"). Include an "escape option" where it seems appropriate (e.g., "Don't know" or "Does not apply"). Use wording and response scales that match the reading ability and knowledge level of the intended respondents. Pretest and revise. Keep the questionnaire as short as possible. Table 2.13 lists a few common question formats and response scales.

**TABLE 2.13**

### Question Format and Response Scale

Type of Question	Example																				
Open-ended	Briefly describe your job goals.																				
Fill-in-the-blank	How many times did you attend formal religious services during the last year? _____ times																				
Check boxes	What is your most common method of communication? <input type="checkbox"/> Cell Phone Call <input type="checkbox"/> Text Message <input type="checkbox"/> Email <input type="checkbox"/> Facebook <input type="checkbox"/> Other																				
Ranked choices	Please evaluate your dining experience: <table style="margin-left: auto; margin-right: auto;"> <tr> <td></td> <td>Excellent</td> <td>Good</td> <td>Fair</td> <td>Poor</td> </tr> <tr> <td>Food</td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> <tr> <td>Service</td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> <tr> <td>Ambiance</td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> </table>		Excellent	Good	Fair	Poor	Food	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Service	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Ambiance	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Excellent	Good	Fair	Poor																	
Food	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																	
Service	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																	
Ambiance	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																	
Pictograms	What do you think of the president's economic policies? (circle one) 																				
Likert scale	Statistics is a difficult subject. <table style="margin-left: auto; margin-right: auto;"> <tr> <td>Strongly Agree</td> <td>Slightly Agree</td> <td>Neither Agree nor Disagree</td> <td>Slightly Disagree</td> <td>Strongly Disagree</td> </tr> <tr> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> </table>	Strongly Agree	Slightly Agree	Neither Agree nor Disagree	Slightly Disagree	Strongly Disagree	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>										
Strongly Agree	Slightly Agree	Neither Agree nor Disagree	Slightly Disagree	Strongly Disagree																	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																	

## Survey Validity and Reliability

Surveys are often called *instruments* because they are thought of as a measurement tool. As a measurement tool, researchers want the survey to be both **valid** and **reliable**. A valid survey is one that measures what the researcher wants to measure. Are the questions worded in such a way that the responses provide information about what the researcher wants to know? A reliable survey is one that is consistent. In other words, over time, will the responses from similar respondents stay the same?

## Question Wording

The way a question is asked has a profound influence on the response. For example, in a *Wall Street Journal* editorial, Fred Barnes tells of a *Reader's Digest* poll that asked two similar questions:

Version 1: I would be disappointed if Congress cut its funding for public television.

Version 2: Cuts in funding for public television are justified to reduce federal spending.

The same 1,031 people were polled in both cases. Version 1 showed 40 percent in favor of cuts, while version 2 showed 52 percent in favor of cuts. The margin of error was  $\pm 3.5$  percent (in “How to Rig a Poll,” June 14, 1995, p. A18). To “rig” the poll, emotional overlays or “loaded” mental images can be attached to the question. In fact, it is often difficult to ask a neutral question without any context. For example:

Version 1: Shall state taxes be cut?

Version 2: Shall state taxes be cut, if it means reducing highway maintenance?

Version 3: Shall state taxes be cut, if it means firing teachers and police?

An unconstrained choice (version 1) makes tax cuts appear to be a “free lunch,” while versions 2 and 3 require the respondent to envision the consequences of a tax cut. An alternative is to use version 1 but then ask the respondent to list the state services that should be cut to balance the budget after the tax cut.

Another problem in wording is to make sure you have covered all the possibilities. For example, how could a widowed independent voter answer questions like these?

Are you married?

Yes

No

What is your party preference?

Democrat

Republican

Avoid overlapping classes or unclear categories. What if the respondent’s father is deceased or is 45 years old?

How old is your father?

35–45

45–55

55–65

65 or older

## Data Quality

Survey responses usually are coded numerically (e.g., 1 = male, 2 = female), although many software packages also can use text variables (nominal data) in certain kinds of statistical tests. Most packages require you to denote missing values by a special character (e.g., blank, period, or asterisk). If too many entries on a given respondent’s questionnaire are flawed or missing, you may decide to discard the entire response.

Other data screening issues include multiple responses (i.e., the respondent chose two responses where one was expected), outrageous replies on fill-in-the-blank questions (e.g., a respondent who claims to work 640 hours a week), “range” answers (e.g., 10–20 cigarettes smoked per day), or inconsistent replies (e.g., a 25-year-old respondent who claims to receive Medicare benefits). Sometimes a follow-up is possible, but in anonymous surveys, you must make the best decisions you can about how to handle anomalous data. Be sure to document your data-coding decisions—not only for the benefit of others but also in case you are asked to explain how you did it (it is easy to forget after a month or two, when you have moved on to other projects).

## Survey Software

Designing and creating a survey is much easier than it used to be. Software is available that automates much of the process, allowing you to use different question formats, skip questions and move to a new section, easily visualize the layout, and other features. Because most surveys

are now administered online, survey software also includes features that allow the respondent to remain anonymous if warranted and prevent respondents from taking the survey twice. One of the most commonly used free applications is SurveyMonkey ([www.surveymonkey.com](http://www.surveymonkey.com)). Qualtrics ([www.qualtrics.com](http://www.qualtrics.com)) offers a 14-day free trial but is used in larger enterprises such as universities. Many other software applications exist that offer similar features to SurveyMonkey or Qualtrics. It is important to remember that survey design, creation, and administration require thoughtful preparation and planning in order to capture the right information and to ensure a high response rate.

## SECTION EXERCISES



- 2.24 What sources of error might you encounter if you want to know (a) about the dating habits of college men, so you go to a dorm meeting and ask students how many dates they have had in the last year; (b) how often people attend religious services, so you stand outside a particular church on Sunday and ask entering individuals how often they attend; (c) how often people eat at McDonald's, so you stand outside a particular McDonald's and ask entering customers how often they eat at McDonald's?
- 2.25 What kind of survey (mail, telephone, interview, web, direct observation) would you recommend for each of the following purposes, and why? What problems might be encountered?
- To estimate the proportion of students at your university who would prefer a web-based statistics class to a regular lecture.
  - To estimate the proportion of students at your university who carry backpacks to class.
  - To estimate the proportion of students at your university who would be interested in taking a two-month summer class in international business with tours of European factories.
  - To estimate the proportion of U.S. business graduates who have taken a class in international business.
- 2.26 What kind of survey (mail, telephone, interview, web, direct observation) would you recommend that a small laundry and dry cleaning business use for each of the following purposes, and why?
- To estimate the proportion of customers preferring opening hours at 7 a.m. instead of 8 a.m.
  - To estimate the proportion of customers who have only laundry and no dry cleaning.
  - To estimate the proportion of residents in the same zip code who spend more than \$20 a month on dry cleaning.
  - To estimate the proportion of its seven employees who think it is too hot inside the building.
- 2.27 What would be the difference in student responses to the two questions shown?
- Version 1: I would prefer that tuition be reduced.
- Version 2: Cuts in tuition are a good idea even if some classes are canceled.
- 2.28 What problems are evident in the wording of these two questions?
- |                                |                                    |
|--------------------------------|------------------------------------|
| What is your race?             | What is your religious preference? |
| <input type="checkbox"/> White | <input type="checkbox"/> Christian |
| <input type="checkbox"/> Black | <input type="checkbox"/> Jewish    |

## Mini Case

## 2.4

### Roles of Colleges

A survey of public opinion on the role of colleges was conducted by *The Chronicle of Higher Education*. Results of the survey showed that 77 percent of respondents agreed it was highly important that colleges prepare its undergraduate students for a career. The percentage of respondents who agreed it was highly important for colleges to prepare students to be responsible citizens was slightly lower, at 67 percent. The survey utilized 1,000 telephone interviews of 20 minutes each, using a random selection of men and women aged 25 through 65. It was conducted on February 25, 2004. The survey was administered by TMR Inc. of Broomall, Pennsylvania. Data were collected and analyzed by GDA Integrated Services, a market research firm in Old Saybrook, Connecticut.

The Likert-type scale labels are weighted toward the positive, which is common when the survey items (roles for colleges in this case) are assumed to be potentially important

and there is little likelihood of a strong negative response. Respondents also were asked for demographic information. Fifty-eight percent were women and 42 percent were men, coming from all states except Alaska and Hawaii. Eleven percent were African American (similar to the national average), but only 6 percent were Hispanic (about 8 percent below the national average). The underrepresentation of Hispanics was due to language barriers, illustrating one difficulty faced by surveys. However, the respondents' incomes, religious affiliations, and political views were similar to the general U.S. population. The random selection method was not specified. Note that firms that specialize in survey sampling generally have access to commercial lists and use their own proprietary methods.

A **data set** consists of all the values of all the variables we have chosen to observe. It often is an array with  $n$  rows and  $m$  columns. Data sets may be **univariate** (one variable), **bivariate** (two variables), or **multivariate** (three or more variables). There are two basic data types: **categorical data** (categories that are described by labels) or **numerical** (meaningful numbers). Numerical data are **discrete** if the values are integers or can be counted or **continuous** if any interval can contain more data values. **Nominal** measurements are names, **ordinal** measurements are ranks, **interval** measurements have meaningful distances between data values, and **ratio** measurements have meaningful ratios and a zero reference point. **Time series** data are observations measured at  $n$  different points in time or over sequential time intervals, while **cross-sectional** data are observations among  $n$  entities such as individuals, firms, or geographic regions. Among **random samples**, **simple random** samples pick items from a list using random numbers, **systematic** samples take every  $k$ th item, **cluster** samples select geographic regions, and **stratified** samples take into account known population proportions. **Nonrandom** samples include convenience or judgment samples, gaining time but sacrificing randomness. **Focus groups** give in-depth information. **Survey design** requires attention to question **wording** and **scale definitions**. **Survey techniques** (mail, telephone, interview, web, direct observation) depend on time, budget, and the nature of the questions and are subject to various sources of error.

binary variable  
bivariate data sets  
categorical data  
census  
cluster sample  
coding  
continuous data  
convenience sampling  
coverage error  
cross-sectional data  
data  
data set  
discrete data  
focus group  
interval data  
interviewer error

judgment sampling  
Likert scale  
measurement error  
multivariate data sets  
nominal data  
nonrandom sampling  
nonresponse bias  
numerical data  
observation  
ordinal data  
parameter  
population  
random numbers  
random sampling  
ratio data  
reliability

response error  
sample  
sampling error  
sampling frame  
sampling with replacement  
sampling without replacement  
selection bias  
simple random sample  
statistics  
stratified sampling  
systematic sampling  
target population  
time series data  
univariate data sets  
validity  
variable

## CHAPTER SUMMARY

## KEY TERMS

## CHAPTER REVIEW

1. Define (a) data, (b) data set, (c) observation, and (d) variable.
2. How do business data differ from scientific experimental data?
3. Distinguish (a) univariate, bivariate, and multivariate data; (b) discrete and continuous data; (c) numerical and categorical data.
4. Define the four measurement levels, and give an example of each.
5. Explain the difference between cross-sectional data and time series data.
6. (a) List three reasons why a census might be preferred to a sample. (b) List three reasons why a sample might be preferred to a census.
7. (a) What is the difference between a parameter and a statistic? (b) What is a target population?

8. (a) List four methods of random sampling. (b) List two methods of nonrandom sampling. (c) Why would we ever use nonrandom sampling? (d) Why is sampling usually done without replacement?
9. List five (a) steps in a survey, (b) issues in survey design, (c) survey types, (d) question scale types, and (e) sources of error in surveys.
10. List advantages and disadvantages of different types of surveys.

## CHAPTER EXERCISES



## DATA TYPES

- 2.29** Which type of data (categorical, discrete numerical, continuous numerical) is each of the following variables?
- a. Age of a randomly chosen tennis player in the Wimbledon tennis tournament.
  - b. Nationality of a randomly chosen tennis player in the Wimbledon tennis tournament.
  - c. Number of double faults in a randomly chosen tennis game at Wimbledon.
- 2.30** Which type of data (categorical, discrete numerical, continuous numerical) is each of the following variables?
- a. Number of spectators at a randomly chosen Wimbledon tennis match.
  - b. Water consumption (liters) by a randomly chosen Wimbledon player during a match.
  - c. Gender of a randomly chosen tennis player in the Wimbledon tennis tournament.
- 2.31** Which measurement level (nominal, ordinal, interval, ratio) is each of the following variables?
- a. A customer's ranking of five new hybrid vehicles.
  - b. Noise level 100 meters from the Dan Ryan Expressway at a randomly chosen moment.
  - c. Number of occupants in a randomly chosen commuter vehicle on the San Diego Freeway.
- 2.32** Which measurement level (nominal, ordinal, interval, ratio) is each of the following variables?
- a. Number of annual office visits by a particular Medicare subscriber.
  - b. Daily caffeine consumption by a six-year-old child.
  - c. Type of vehicle driven by a college student.
- 2.33** Below are five questions from a survey of MBA students. Answers were written in the blank at the left of each question. For each question, state the data type (categorical, discrete numerical, or continuous numerical) and measurement level (nominal, ordinal, interval, ratio). Explain your reasoning. If there is doubt, discuss the alternatives.
- |          |  |
|----------|--|
| _____ Q1 | What is your gender? (Male = 0, Female = 1)  |
| _____ Q2 | What is your approximate undergraduate college GPA? (1.0 to 4.0)   |
| _____ Q3 | About how many hours per week do you expect to work at an outside job this semester?                               |
| _____ Q4 | What do you think is the ideal number of children for a married couple?  |
| _____ Q5 | On a 1 to 5 scale, which best describes your parents?<br>1 = Mother clearly dominant ↔ 5 = Father clearly dominant |
- 2.34** Below are five questions from a survey of MBA students. Answers were written in the blank at the left of each question. For each question, state the data type (categorical, discrete numerical, or continuous numerical) and measurement level (nominal, ordinal, interval, ratio). Explain your reasoning. If there is doubt, discuss the alternatives.
- |           |   |
|-----------|---|
| _____ Q6  | On a 1 to 5 scale, assess the current job market for your undergraduate major. 1 = Very bad ↔ 5 = Very good |
| _____ Q7  | During the last month, how many times has your schedule been disrupted by car trouble?                      |
| _____ Q8  | About how many years of college does the more-educated one of your parents have? (years)                    |
| _____ Q9  | During the last year, how many traffic tickets (excluding parking) have you received?                       |
| _____ Q10 | Which political orientation most nearly fits you? (1 = Liberal, 2 = Middle-of-Road, 3 = Conservative)       |
- 2.35** Below are five questions from a survey of MBA students. Answers were written in the blank at the left of each question. For each question, state the data type (categorical, discrete numerical, or continuous numerical) and measurement level (nominal, ordinal, interval, ratio). Explain your reasoning. If there is doubt, discuss the alternatives.
- |           |   |
|-----------|---|
| _____ Q11 | What is the age of the car you usually drive? (years)                           |
| _____ Q12 | About how many times in the past year did you attend formal religious services? |



- \_\_\_\_\_ Q13 How often do you read a daily newspaper? (0 = Never, 1 = Occasionally, 2 = Regularly)
- \_\_\_\_\_ Q14 Can you conduct simple transactions in a language other than English? (0 = No, 1 = Yes)
- \_\_\_\_\_ Q15 How often do you exercise (aerobics, running, etc)? (0 = Not at All, 1 = Sometimes, 2 = Regularly)

**2.36** Identify the following data as either time series or cross-sectional.

- The 2017 CEO compensation of the 500 largest U.S. companies.
- The annual compensation for the CEO of Coca-Cola Enterprises from 2010 to 2017.
- The weekly revenue for a Noodles & Company restaurant for the 52 weeks in 2017.
- The number of skiers on the mountain on Christmas Day 2017 at each of the ski mountains owned by Vail Resorts.

**2.37** Identify the following data as either time series or cross-sectional.

- The number of rooms booked each night for the month of January 2017 at a Vail Resorts hotel.
- The amount spent on books at the start of this semester by each student in your statistics class.
- The number of Caesar salads sold for the week of April 19, 2017, at each Noodles & Company restaurant.
- The stock price of Coca-Cola Enterprises on May 1st for each of the last 10 years.

### **SAMPLING METHODS**

**2.38** Would you use a sample or a census to measure each of the following? Why? If you are uncertain, explain the issues.

- The number of cans of Campbell's soup on your local supermarket's shelf today at 6:00 p.m.
- The proportion of soup sales last week in Boston that was sold under the Campbell's brand.
- The proportion of Campbell's brand soup cans in your family's pantry.

**2.39** Would you use a sample or census to measure each of the following?

- The number of workers currently employed by Campbell Soup Company.
- The average price of a can of Campbell's Cream of Mushroom soup.
- The total earnings of workers employed by Campbell Soup Company last year.

**2.40** Is each of the following a parameter or a statistic? If you are uncertain, explain the issues.

- The number of cans of Campbell's soup sold last week at your local supermarket.
- The proportion of all soup in the United States that was sold under the Campbell's brand last year.
- The proportion of Campbell's brand soup cans in the family pantries of 10 students.

**2.41** Is each of the following a parameter or statistic?

- The number of visits to a pediatrician's office last week.
- The number of copies of John Grisham's most recent novel sold to date.
- The total revenue realized from sales of John Grisham's most recent novel.

**2.42** Recently, researchers estimated that 76.8 percent of global e-mail traffic was spam. Could a census be used to update this estimate? Why or why not?

**2.43** A certain health maintenance organization (HMO) is studying its daily office routine. It collects information on three variables: the number of patients who visit during a day, the patient's complaint, and the waiting time until each patient sees a doctor. (a) Which variable is categorical? (b) Identify the two quantitative variables, and state whether they are discrete or continuous.

**2.44** There are 327 official ports of entry in the United States. The Department of Homeland Security selects 15 ports of entry at random to be audited for compliance with screening procedures of incoming travelers through the primary and secondary vehicle and pedestrian lanes. What kind of sample is this (simple random, systematic, stratified, cluster)?

**2.45** The IRS estimates that the average taxpayer spent 3.7 hours preparing Form 1040 to file a tax return. Could a census be used to update this estimate for the most recent tax year? Why or why not?

**2.46** The General Accounting Office conducted random testing of retail gasoline pumps in Michigan, Missouri, Oregon, and Tennessee. The study concluded that 49 percent of gasoline pumps nationwide are mislabeled by more than one-half of an octane point. What kind of sampling technique was most likely to have been used in this study?




- 2.47 Arsenic (a naturally occurring, poisonous metal) in home water wells is a common threat. (a) What sampling method would you use to estimate the arsenic levels in wells in a rural county to see whether the samples violate the EPA limit of 10 parts per billion (ppb)? (b) Is a census possible?
- 2.48 Would you expect Starbucks to use a sample or census to measure each of the following? Explain.
- The percentage of repeat customers at a certain Starbucks on Saturday mornings.
  - The number of chai tea latte orders last Saturday at a certain Starbucks.
  - The average temperature of Starbucks coffee served on Saturday mornings.
  - The revenue from coffee sales as a percentage of Starbucks' total revenue last year.
- 2.49 Would you expect Noodles & Company to use a sample or census to measure each of the following? Explain.
- The annual average weekly revenue of each Noodles restaurant.
  - The average number of weekly lunch visits by customers.
  - The customer satisfaction rating of a new dessert.
  - The number of weeks in a year that a restaurant sells more bottled beverages than fountain drinks.
- 2.50 A financial magazine publishes an annual list of major stock funds. Last year, the list contained 1,699 funds. What method would you recommend to obtain a sample of 20 stock funds?
- 2.51 Examine each of the following statistics. Which sampling method was most likely to have been used (simple random, systematic, stratified, cluster)?
- A survey showed that 30 percent of U.S. businesses have fired an employee for inappropriate web surfing, such as gambling, watching porn, or shopping.
  - Surveyed doctors report that 59 percent of patients do not follow their prescribed treatment.
  - The Internal Revenue Service reports that, based on a sample of individual taxpayers, 80 percent of those who failed to pay what they owed did so through honest errors or misinterpretation of the tax code.
  - In Spain, per capita consumption of cigarettes is 1,265 compared with 1,083 in the United States.
- 2.52 The National Claims History (NCH) contains records for 999,645 Medicare patients who were discharged from acute care hospitals in October 2008. The Department of Health and Human Services performed a detailed audit of adverse medical events on a random sample of 780 drawn at random without replacement by assigning a random number to each patient on the list and then choosing random integers between 1 and 999,645. (a) What kind of sample is this (random, systematic, stratified, cluster)? (b) Is this population effectively infinite?
- 2.53 Prior to starting a recycling program, a city decides to measure the quantity of garbage produced by single-family homes in various neighborhoods. This experiment will require weighing garbage on the day it is set out. (a) What sampling method would you recommend, and why? (b) What would be a potential source of sampling error?
- 2.54 A university wanted to survey alumni about their interest in lifelong learning classes. They mailed questionnaires to a random sample of 600 alumni from their database of over 30,000 recent graduates. Would you consider this population to be effectively infinite?
- 2.55 The U.S. Fisheries and Wildlife Service requires that scallops harvested from the ocean must weigh at least  $1/36$  pound. The harbormaster at a Massachusetts port randomly selected 18 bags of scallops from 11,000 bags on an arriving vessel. The average scallop weight from the 18 bags was  $1/39$  pound. (a) Would the population of 11,000 bags be considered effectively infinite in this case? (b) Which value represents a sample statistic:  $1/36$  or  $1/39$ ? (See *Interfaces* 25, no. 2 [March–April 1995], p. 18.)
- 2.56 A marketing research group wanted to collect information from existing and potential customers on the appeal of a new product. They sent out surveys to a random sample of 1,200 people from their database of over 25,000 current and potential customers. Would you consider this population to be effectively infinite?
- 2.57 Households can sign up for a telemarketing “no-call list.” How might households who sign up differ from those who don’t? What biases might this create for telemarketers promoting (a) financial planning services, (b) carpet cleaning services, and (c) vacation travel packages?

**SURVEYS AND SCALES**

- 2.58** Suggest response check boxes for these questions. In each case, what difficulties do you encounter as you try to think of appropriate check boxes?
- Where are you employed?
  - What is the biggest issue facing the next U.S. president?
  - Are you happy?
- 2.59** Suggest both a Likert scale question and a response scale to measure the following:
- A student's rating of a particular statistics professor.
  - A voter's satisfaction with the president's economic policy.
  - An HMO patient's perception of waiting time to see a doctor.
- 2.60** What level of measurement (nominal, ordinal, interval, ratio) is appropriate for the movie rating system that you see in *TV Guide* (★, ★★, ★★★, ★★★★)? Explain your reasoning.
- 2.61** Insurance companies are rated by several rating agencies. The Fitch 20-point scale is AAA, AA+, AA, AA-, A+, A, A-, BBB+, BBB, BBB-, BB+, BB, BB-, B+, B, B-, CCC+, CCC, CCC-, DD. (a) What level of measurement does this scale use? (b) To assume that the scale uses interval measurements, what assumption is required?
- 2.62** A tabletop survey by a restaurant asked the question shown below. (a) What kind of response scale is this? (b) Suggest an alternative response scale that would be more sensitive to differences in opinion. (c) Suggest possible sources of bias in this type of survey.
- Were the food and beverage presentations appealing?
- Yes       No

**MINI PROJECTS**


- 2.63** Give *two* original examples of (a) discrete data and (b) continuous data. In each example, explain and identify any ambiguities that might exist. *Hint:* Consider data describing your own life (e.g., your sports performance or financial or academic data). You need *not* list all the data; merely describe them and show a few typical data values.
- 2.64** Give *two* original examples of (a) time series data and (b) cross-sectional data. *Hint:* Do not restrict yourself to published data. You need *not* list all the data; merely describe them and show a few typical data values.
- 2.65** Devise a practical sampling method to collect data to estimate each of the following parameters.
- Percentage of an HMO's patients who make more than five office visits per year.
  - Noise level (measured in decibels) in neighborhoods 100 meters from a certain freeway.
  - Percentage of bank mortgages issued to first-time borrowers last year.
- 2.66** Devise a practical sampling method to collect data to estimate each of the following parameters.
- Percentage of peanuts in a can of Planter's Mixed Nuts.
  - Average price of gasoline in your area.
  - Average flight departure delay for Southwest Airlines in Salt Lake City.
- 2.67** Below are 64 names of employees at NilCo. Colors denote different departments (finance, marketing, purchasing, engineering). Sample eight names from the display shown by using (a) simple random sampling, (b) systematic sampling, and (c) cluster sampling. Try to ensure that every name has an equal chance of being picked. Which sampling method seems most appropriate?

 **PickEight**


Floyd	Sid	LaDonna	Tom	Mabel	Nicholas	Bonnie	Deepak
Nathan	Ginnie	Mario	Claudia	Dmitri	Kevin	Blythe	Dave
Lou	Tim	Peter	Jean	Mike	Jeremy	Chad	Doug
Loretta	Erik	Jackie	Juanita	Molly	Carl	Buck	Janet
Anne	Joel	Moira	Marnie	Ted	Greg	Duane	Amanda
Don	Gadis	Balaji	Al	Takisha	Dan	Ryan	Sam
Graham	Scott	Lorin	Vince	Jody	Brian	Tania	Ralph
Bernie	Karen	Ed	Liz	Erika	Marge	Gene	Pam

- 2.68 From the display below, pick five cards (without replacement) by using random numbers. Explain your method. Why would the other sampling methods not work well in this case?

A ♠	A ♥	A ♣	A ♦
K ♠	K ♥	K ♣	K ♦
Q ♠	Q ♥	Q ♣	Q ♦
J ♠	J ♥	J ♣	J ♦
10 ♠	10 ♥	10 ♣	10 ♦
9 ♠	9 ♥	9 ♣	9 ♦
8 ♠	8 ♥	8 ♣	8 ♦
7 ♠	7 ♥	7 ♣	7 ♦
6 ♠	6 ♥	6 ♣	6 ♦
5 ♠	5 ♥	5 ♣	5 ♦
4 ♠	4 ♥	4 ♣	4 ♦
3 ♠	3 ♥	3 ♣	3 ♦
2 ♠	2 ♥	2 ♣	2 ♦

- 2.69 Photocopy the exhibit below (omit these instructions) and show it to a friend or classmate. Ask him/her to choose a number at random and write it on a piece of paper. Collect the paper. Repeat for *at least* 20 friends/classmates. Tabulate the results. Were all the numbers chosen equally often? If not, which were favored or avoided? Why?  **PickOne**

0	11	17	22
8	36	14	18
19	28	6	41
12	3	5	0

- 2.70 Ask each of 20 friends or classmates to choose a whole number between 1 and 5. Tabulate the results. Do the results seem random? If not, can you think of any reasons?
- 2.71 You can test Excel's algorithm for selecting random integers with a simple experiment. Enter =RANDBETWEEN(1,2) into cell A1 and then copy it to cells A1:E20. This creates a data block of 100 cells containing either a one or a two. In cell G1 type =COUNTIF(A1:E20,"=1") and in cell G2 type =COUNTIF(A1:E20,"=2"). Highlight cells G1 and G2 and create a column chart. Click on the vertical axis scale, and set the lower limit to 0 and upper limit to 100. Then hold down the F9 key and observe the chart. Are you convinced that, on average, you are getting about 50 ones and 50 twos? *Ambitious Students*: Generalize this experiment to integers 1 through 5.  **RandBetween**

## RELATED READING

### Guides to Data Sources

Clayton, Gary E., and Martin Giesbrecht. *A Guide to Everyday Economic Statistics*. 7th ed. McGraw-Hill, 2009.

### Sampling and Surveys

Best, Joel. *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists*. University of California Press, 2012.

Cooper, Donald R., and Pamela S. Schindler. *Business Research Methods*. 12th ed. McGraw-Hill, 2014.

Fowler, Floyd J. *Survey Research Methods*. 4th ed. Sage, 2009.

Groves, Robert M., et al. *Survey Methodology*. 2nd ed. Wiley, 2009.

Johnson, John H., and Mike Gluck. *Everydata: The Misinformation Hidden in the Little Data You Consume Every Day*. Bibliomotion Inc., 2016.

Levy, Paul S., and Stanley Lemeshow. *Sampling of Populations*. 4th ed. Wiley, 2008.

Lohr, Sharon L. *Sampling: Design & Analysis*. 2nd ed. Cengage, 2010.

Lyberg, Lars, and Paul Blemer. *Introduction to Survey Quality*. Wiley Europe, 2003.

Mathieson, Kieran, and David P. Doane. "Using Fine-Grained Likert Scales in Web Surveys." *Alliance Journal of Business Research* 1, no. 1 (2006), pp. 27-34.









Scheaffer, Richard L.; William Mendenhall; and R. Lyman Ott. *Elementary Survey Sampling*. 7th ed. Brooks/Cole, 2012.


Thompson, Steven K. *Sampling*. 3rd ed. Wiley, 2012.

## CHAPTER 2 More Learning Resources

You can access these *LearningStats* demonstrations through McGraw-Hill's Connect® to help you understand random sampling.



Topic	LearningStats Demonstrations
Sampling	<ul style="list-style-type: none"> <li> Sampling Methods</li> <li> Who Gets Picked?</li> <li> Randomizing a List</li> <li> Pick a Card</li> <li> Excel's RANDBETWEEN Function</li> </ul>
Data sources	<ul style="list-style-type: none"> <li> Web Data Sources</li> <li> Survey Tips</li> <li> Sampling Plans</li> </ul>

Key:  = Excel  = PowerPoint