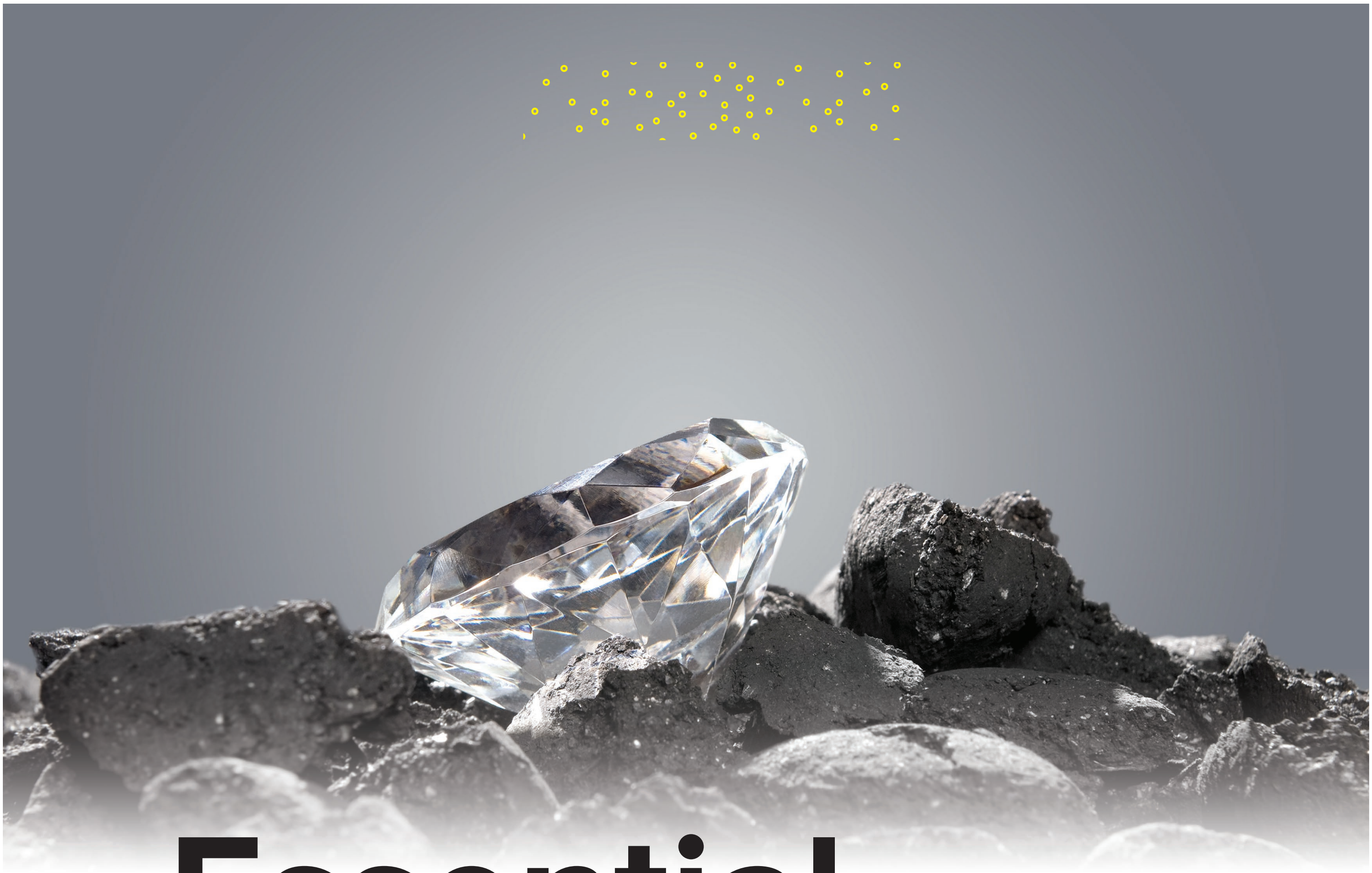
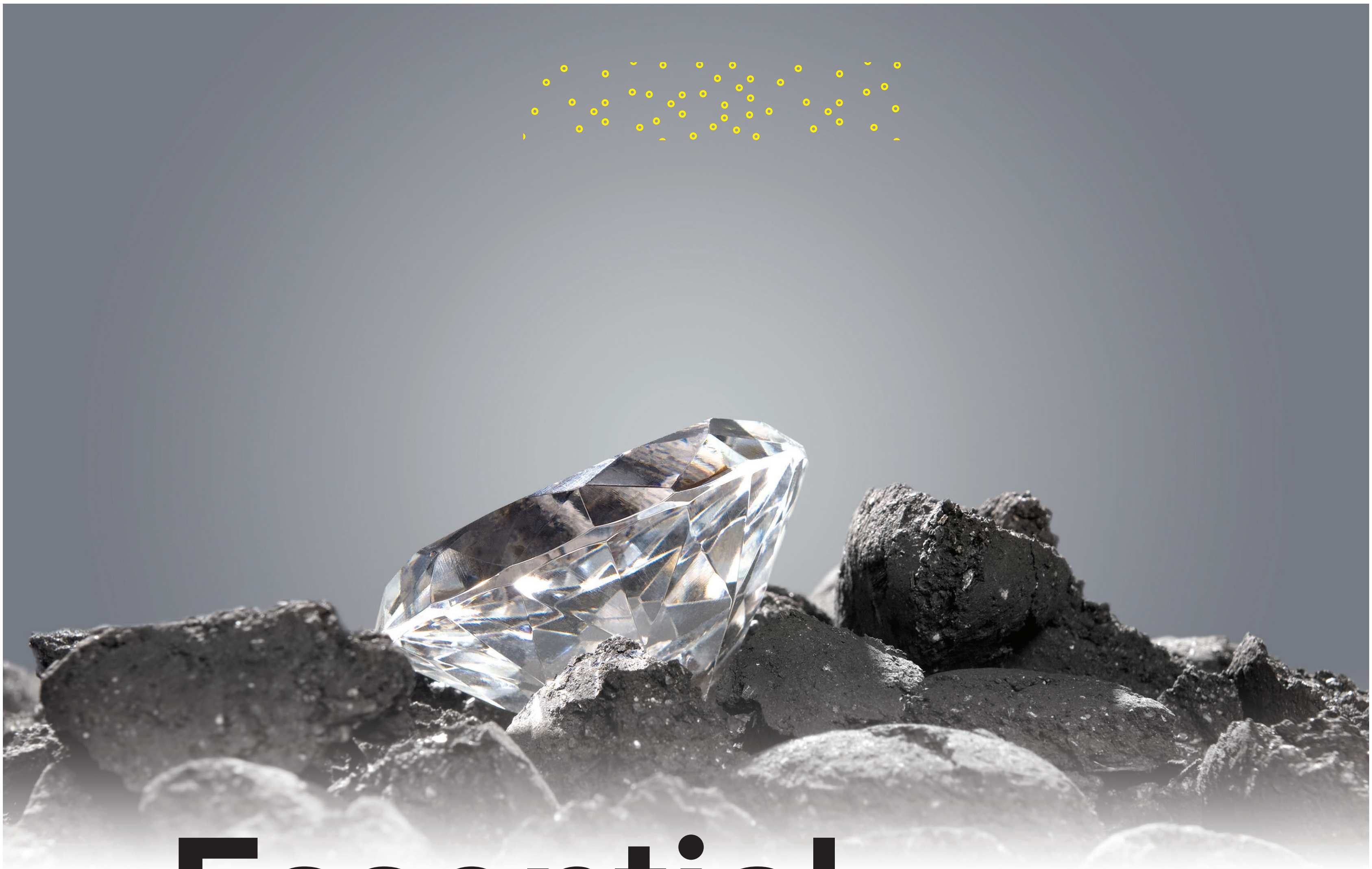


Essential STATISTICS



Essential STATISTICS

Third Edition

William Navidi
Colorado School of Mines

Barry Monk
Middle Georgia State University

Mc Graw Hill

ESSENTIAL STATISTICS, THIRD EDITION

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 LWI 26 25 24 23 22 21

ISBN 978-1-260-35979-4 (bound edition)
MHID 1-260-35979-4 (bound edition)

ISBN 978-1-260-49216-3 (loose-leaf edition)
MHID 1-260-49216-8 (loose-leaf edition)

ISBN 978-1-264-57183-3 (instructor's edition)
MHID 1-264-57183-6 (instructor's edition)

Portfolio Manager: *Chrissy Skogen*
Product Developer: *Megan Platt*
Marketing Manager: *Emily Digiovanna*
Content Project Managers: *Jeni McAtee, Rachael Hillebrand*
Buyer: *Laura Fuller*
Designer: *Beth Blech*
Content Licensing Specialist: *Jacob Sullivan*
Cover Image: *Joe Belanger/Alamy*
Compositor: *Aptara®, Inc.*

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: Navidi, William Cyrus, author. | Monk, Barry (Barry J.), author.
Title: Essential statistics / William Navidi, Colorado School of Mines, Barry Monk, Middle Georgia State University.
Description: Third edition. | New York, NY : McGraw-Hill Education, [2022] | Includes index.
Identifiers: LCCN 2020039533 (print) | LCCN 2020039534 (ebook) | ISBN 9781260359794 (hardcover) | ISBN 9781260492163 (spiral bound) | ISBN 9781264571833 (hardcover) | ISBN 9781260492194 (ebook) | ISBN 9781260492293 (ebook other)
Subjects: LCSH: Mathematical statistics—Textbooks.
Classification: LCC QA276.12 .N386 2022 (print) | LCC QA276.12 (ebook) | DDC 519.5—dc23
LC record available at https://lccn.loc.gov/2020039533
LC ebook record available at https://lccn.loc.gov/2020039534

mheducation.com/highered

To Catherine, Sarah, and Thomas

—William Navidi

To Shaun, Dawn, and Ben

—Barry Monk

About the Authors

©William Navidi

William Navidi is a professor of Applied Mathematics and Statistics at the Colorado School of Mines in Golden, Colorado. He received a Bachelor's degree in Mathematics from New College, a Master's degree in Mathematics from Michigan State University, and a Ph.D. in Statistics from the University of California at Berkeley. Bill began his teaching career at the County College of Morris, a two-year college in Dover, New Jersey. He has taught mathematics and statistics at all levels, from developmental through the graduate level. Bill has written two Engineering Statistics textbooks for McGraw Hill. In his spare time, he likes to play racquetball.

©Dawn Sherry

Barry Monk is a Professor of Mathematics with Middle Georgia State University in Macon, Georgia. Barry received a Bachelor of Science in Mathematical Statistics, a Master of Arts in Mathematics specializing in Optimization and Statistics, and a Ph.D. in Applied Mathematics, all from the University of Alabama. Barry has been teaching Introductory Statistics since 1992 in the classroom and online environments. Barry has a minor in Creative Writing and is a skilled jazz pianist.

Brief Contents

Contents

Preface

This book is designed for an introductory course in statistics. In addition to presenting the mechanics of the subject, we have endeavored to explain the concepts behind them in a writing style as straightforward, clear, and engaging as we could make it. As practicing statisticians, we have done everything possible to ensure that the material is accurate and correct. We believe that this book will enable instructors to explore statistical concepts in depth yet remain easy for students to read and understand.

To achieve this goal, we have incorporated a number of useful pedagogical features:

Features

- **Check Your Understanding Exercises:** After each concept is explained, one or more exercises are immediately provided for students to be sure they are following the material. These exercises provide students with confidence that they are ready to go on, or alert them to the need to review the material just covered.
- **Explain It Again:** Many important concepts are reinforced with additional explanation in these marginal notes.
- **Real Data:** Statistics instructors universally agree that the use of real data engages students and convinces them of the usefulness of the subject. A great many of the examples and exercises use real data. Some data sets explore topics in health or social sciences, while others are based in popular culture such as movies, contemporary music, or video games.
- **Integration of Technology:** Many examples contain screenshots from the TI-84 Plus calculator, MINITAB, and Excel. Each section contains detailed, step-by-step instructions, where applicable, explaining how to use these forms of technology to carry out the procedures explained in the text.
- **Interpreting Technology:** Many exercises present output from technology and require the student to interpret the results.
- **Write About It:** These exercises, found at the end of each chapter, require students to explain statistical concepts in their own words.
- **Case Studies:** Each chapter begins with a discussion of a real problem. At the end of the chapter, a case study demonstrates applications of chapter concepts to the problem.
- **In-Class Activities:** At the end of each chapter, activities are suggested that reinforce some concepts presented in the chapter.

Flexibility

We have endeavored to make our book flexible enough to work effectively with a wide variety of instructor styles and preferences. We cover both the *P*-value and critical value approaches to hypothesis testing, so instructors can choose to cover either or both of these methods.

Instructors differ in their preferences regarding the depth of coverage of probability. A light treatment of the subject may be obtained by covering Section 4.1 and skipping the rest of the chapter. More depth can be obtained by covering Section 4.2.

Supplements

Supplements, including a Corequisite Workbook, online homework, videos, guided student notes, and PowerPoint presentations, play an increasingly important role in the educational process. As authors, we have adopted a hands-on approach to the development of our supplements, to make sure that they are consistent with the style of the text and that they work effectively with a variety of instructor preferences. In particular, our online homework package offers instructors the flexibility to choose whether the solutions that students view are based on tables or technology, where applicable.

New in This Edition

The third edition of the book is intended to extend the strengths of the second. Some of the changes are:

- Discussions of the investigative process of statistics have been added, in accordance with recommendations of the GAISE report.
- In-class activities have been added to each chapter
- Material on the ratio and interval levels of measurement has been added.
- Material on bell-shaped histograms has been added.

- A discussion of the use of sample means to estimate population means has been added.
- Material on the uniform distribution has been added.
- A new section on multiple testing has been added.
- A new objective on the reasoning used in hypothesis testing has been added.
- New conceptual exercises regarding assumptions in constructing confidence intervals and performing hypothesis tests have been added.
- Additional material on Type I and Type II errors has been added.
- Objectives on the relationship between confidence intervals and the margin of error, calculating the sample size needed for a confidence interval of a given width, and the difference between confidence and probability are now presented in a context where the population standard deviation is unknown.
- Objectives on the relationship between confidence intervals and hypothesis tests, the relationship between the level of a test and the probability of error, the importance of reporting *P*-values, and the difference between statistical and practical significance are now presented in a context where the population standard deviation is unknown.
- Material on confidence intervals and hypothesis tests for paired samples now immediately follows the corresponding material for independent samples.
- A large number of new exercises have been included, many of which involve real data from recent sources.
- A large number of new exercises have been added to the online homework system, ALEKS. These include new conceptual questions and stepped-out solutions for the TI-84 Plus calculator and Excel.
- Several of the case studies have been updated.
- The exposition has been improved in a number of places.

William Navidi
Barry Monk

Acknowledgments

We are indebted to many people for contributions at every stage of development. Colleagues and students who reviewed the evolving manuscript provided many valuable suggestions. In particular, Charla Baker, John Trimboli, Don Brown, and Duane Day contributed to the supplements, and Mary Wolfe helped create the video presentations. Ashlyn Munson contributed a number of exercises, Abby Noble, Lindsay Lewis, and Ryan Melendez contributed many valuable ideas for revisions, and Tim Chappell and Lisa Rombes played an important role in the development of our digital content.

The staff at McGraw Hill has been extremely capable and supportive. Project Manager Jeni McAtee was always patient and helpful. We owe a debt of thanks to Emily DiGiovanna, Debbie McFarland, and Mary Ellen Rahn for their creative marketing and diligence in spreading the word about our book. We appreciate the guidance of our editors, Chrissy Skogen and Megan Platt, whose input has considerably improved the final product.

William Navidi
Barry Monk

Feedback from Statistics Instructors

Paramount to the development of ***Essential Statistics*** was the invaluable feedback provided by the instructors from around the country who reviewed the manuscript while it was in development.

A Special Thanks to All of the Symposia and Focus Group Attendees

James Adair, *Dyersburg State Community College*
Andrea Adlman, *Ventura College*
Leandro Alvarez, *Miami Dade College*
Simon Aman, *City Colleges of Chicago*
Diane Benner, *Harrisburg Area Community College*
Karen Brady, *Columbus State Community College*
Liliana Brand, *Northern Essex Community College*
Denise Brown, *Collin College–Spring Creek*
Don Brown, *Middle Georgia State University*
Mary Brown, *Harrisburg Area Community College*
Gerald Busald, *San Antonio College*
Anna Butler, *Polk State College*
Robert Cappetta, *College of DuPage*
Joe Castillo, *Broward College*
Michele Catterton, *Harford Community College*
Tim Chappell, *Metropolitan Communiity College - Penn Valley*
Ivette Chuca, *El Paso Community College*
James Coker, *University of Maryland Global Campus*
James Condor, *State College of Florida*
Milena Cuellar, *LaGuardia Community College*
Phyllis Curtiss, *Grand Valley State University*
Hema Deshmukh, *Mercyhurst University*
Mitra Devkota, *Shawnee State University*
Sue Jones Dobbyn, *Pellissippi State Community College*
Nataliya Doroshenko, *University of Memphis*
Rob Eby, *Blinn College–Bryan Campus*
Charles Wayne Ehler, *Anne Arundel Community College*
Franco Fedele, *University of West Florida*
Robert Fusco, *Broward College*
Wojciech Golik, *Lindenwood University*
Tim Grant, *Southwestern Illinois College*
Nina Greenberg, *University of New Mexico*
Todd Hendricks, *Georgia State University, Perimeter College*
Mary Hill, *College of DuPage*
Steward Huang, *University of Arkansas–Fort Smith*
Vera Hu-Hyneman, *Suffolk County Community College*
Laura Iossi, *Broward College*
Vera Ioudina, *Texas State University*
Brittany Juraszek, *Santa Fe College*
Maryann Justinger, *Erie Community College–South Campus*
Joseph Karnowski, *Norwalk Community College*
Esmarie Kennedy, *San Antonio College*
Lynette Kenyon, *Collin College–Plano*
Raja Khoury, *Collin College–Plano*
Alexander Kolesnik, *Ventura College*
Holly Kresch, *Diablo Valley College*
JoAnn Kump, *West Chester University*
Dan Kumpf, *Ventura College*
Erica Kwiatkowski-Egizio, *Joliet Junior College*
Hui Liu, *Georgia State University*
Pam Lowry, *Bellevue College*
Corey Manchester, *Grossmont College*
Scott McDaniel, *Middle Tennessee State University*
Mikal McDowell, *Cedar Valley College*
Ryan Melendez, *Arizona State University*
Lynette Meslinsky, *Erie Community College*

Penny Morris, *Polk State College*
Brittany Mosby, *Pellissippi State Community College*
Cindy Moss, *Skyline College*
Kris Mudunuri, *Long Beach City College*
Linda Myers, *Harrisburg Area Community College*
Sean Nguyen, *San Francisco State University*
Kim Page, *Middle Tennessee State University*
Ronald Palcic, *Johnson County Community College*
Matthew Pragel, *Harrisburg Area Community College*
Blanche Presley, *Middle Georgia State University*
Ahmed Rashed, *Richland College*
Cyndi Roemer, *Union County College*
Ginger Rowell, *Middle Tennessee State University*
Sudipta Roy, *Kanakee Community College*
Douglas Ruvolo, *University of Akron*
Ligo Samuel, *Austin Peay State University*
Jamal Salahat, *Owens State Community College*
Kathy Shay, *Middlesex County College*
Laura Shick, *Clemson University*
Larry Shrewsbury, *Southern Oregon University*
Shannon Solis, *San Jacinto College-North*
Tommy Thompson, *Cedar Valley College*
John Trimboli, *Middle Georgia State University*
Rita Sowell, *Volunteer State Community College*
Chris Turner, *Pensacola State College*
Jo Tucker, *Tarrant County College*
Dave Vinson, *Pellissippi State Community College*
Laurie Waites, *University of South Alabama*
Henry Wakhungu, *Indiana University*
Bin Wang, *University of South Alabama*
Daniel Wang, *Central Michigan University*
Rachel Webb, *Portland State University*
Jennifer Zeigenfuse, *Anne Arundel Community College*

Manuscript Review Panels

Alisher Abdullayev, *American River College*
Andrea Adlman, *Ventura College*
Olcay Akman, *Illinois State University*
Raid Amin, *University of West Florida*
Wesley Anderson, *Northwest Vista College*
Peter Arvanites, *Rockland Community College*
Diana Asmus, *Greenville Technical College*
John Avioli, *Christopher Newport University*
Robert Bass, *Gardner-Webb University*
Robbin Bates-Yelverton, *Park University*
Lynn Beckett-Lemus, *El Camino College*
Diane Benner, *Harrisburg Area Community College*
Abraham Biggs, *Broward College*
Wes Black, *Illinois Valley Community College*
Gregory Bloxom, *Pensacola State College*
Dale Bowman, *University of Memphis*
Brian Bradie, *Christopher Newport University*
Tonia Broome, *Gaston College*
Donna Brouilette, *Georgia State University, Perimeter College*
Allen Brown, *Wabash Valley College*
Denise Brown, *Collin Community College*
Don Brown, *Middle Georgia State University*
Mary Brown, *Harrisburg Area Community College*
Jennifer Bryan, *Oklahoma Christian University*
William Burgin, *Gaston College*
Gerald Busald, *San Antonio College*
David Busekist, *Southeastern Louisiana University*
Lynn Cade, *Pensacola State College*
Elizabeth Carrico, *Illinois Central College*
Connie Carroll, *Guilford Technical Community College*
Joseph Castillo, *Broward College*
Linda Chan, *Mount San Antonio College & Pasadena City College*
Ayona Chatterjee, *University of West Georgia*
Chand Chauhan, *Indiana University Purdue University Fort Wayne*
Pinyuen Chen, *Syracuse University*
Askar Choudhury, *Illinois State University*
Lee Clendenning, *University of North Georgia*
James Condor, *State College of Florida–Manatee*
Natalie Creed, *Gaston College*
John Curran, *Eastern Michigan University*
John Daniels, *Central Michigan University*
Shibasish Dasgupta, *University of South Alabama*
Nataliya Doroshenko, *University of Memphis*
Brandon Doughery, *Montgomery County Community College*
Larry Dumais, *American River College*
Christina Dwyer, *State College of Florida–Manatee*
Wayne Ehler, *Anne Arundel Community College*
Mark Ellis, *Central Piedmont Community College*
Angela Everett, *Chattanooga State Technical Community College*
Franco Fedele, *University of West Florida*
Harshini Fernando, *Purdue University—North Central*
Art Fortgang, *Southern Oregon University*
Thomas Fox, *Cleveland State Community College*
Robert Fusco, *Broward College*
Linda Galloway, *Kennesaw State University*
David Garth, *Truman State University*
Sharon Giles, *Grossmont Community College*
Mary Elizabeth Gore, *Community College of Baltimore County*
Carrie Grant, *Flagler College*

Delbert Greear, *University of North Georgia*
Jason Greshman, *Nova Southeastern University*
David Gurney, *Southeastern Louisiana University*
Chris Hail, *Union University–Jackson*
Ryan Harper, *Spartanburg Community College*
Phillip Harris, *Illinois Central College*
James Harrington, *Adirondack Community College*
Matthew He, *Nova Southeastern University*
Mary Beth Headlee, *State College of Florida–Manatee*
James Helmreich, *Marist College*
Todd Hendricks, *Georgia State University, Perimeter College*
Jada Hill, *Richland College*
Mary Hill, *College of DuPage*
William Huepenbecker, *Bowling Green State University–Firelands*
Patricia Humphrey, *Georgia Southern University*
Nancy Johnson, *State College of Florida–Manatee*
Maryann Justinger, *Erie Community College–South Campus*
Joseph Karnowski, *Norwalk Community College*
Susitha Karunaratne, *Purdue University—North Central*
Ryan Kasha, *Valencia College—West Campus*
Joseph Kazimir, *East Los Angeles College*
Esmarie Kennedy, *San Antonio College*
Lynette Kenyon, *Collin College*
Gary Kersting, *North Central Michigan College*
Raja Khoury, *Collin College*
Heidi Kiley, *Suffolk County Community College–Selden*
Daniel Kim, *Southern Oregon University*
Ann Kirkpatrick, *Southeastern Louisiana University*
John Klages, *County College of Morris*
Karon Klipple, *San Diego City College*
Matthew Knowlen, *Horry Georgetown Tech College*
JoAnn Kump, *West Chester University*
Alex Kolesnik, *Ventura College*
Bohdan Kunciw, *Salisbury University*
Erica Kwiatkowski-Egizio, *Joliet Junior College*
William Langston, *Finger Lakes Community College*
Tracy Leshan, *Baltimore City Community College*
Nicole Lewis, *East Tennessee State University*
Jiawei Liu, *Georgia State University*
Fujia Lu, *Endicott College*
Timothy Maharry, *Northwestern Oklahoma State University*
Aldo Maldonado, *Park University*
Kenneth Mann, *Catawba Valley Community College*
James Martin, *Christopher Newport University*
Erin Martin-Wilding, *Parkland College*
Amina Mathias, *Cecil College*
Catherine Matos, *Clayton State University*
Angie Matthews, *Broward College*
Mark McFadden, *Montgomery County Community College*
Karen McKarnin, *Allen Community College*
Penny Morris, *Polk Community College*
B. K. Mudunuri, *Long Beach City College–CalPoly Pomona*
Linda Myers, *Harrisburg Area Community College*
Miroslaw Mystkowski, *Gardner-Webb University*
Shai Neumann, *Brevard Community College*
Francis Nkansah, *Bunker Hill Community College*
Karen Orr, *Roane State Community College*
Richard Owens, *Park University*
Irene Palacios, *Grossmont College*
Luca Petrelli, *Mount Saint Mary's University*
Blanche Presley, *Middle Georgia State University*
Robert Prince, *Berry College*
Richard Puscas, *Georgia State University, Perimeter College*
Ramaswamy Radhakrishnan, *Illinois State University*
Leela Rakesh, *Central Michigan University*
Gina Reed, *University of North Georgia*
Andrea Reese, *Daytona State College–Daytona Beach*
Jim Robison-Cox, *Montana State University*
Alex Rolon, *Northampton Community College*
Jason Rosenberry, *Harrisburg Area Community College*
Yolanda Rush, *Illinois Central College*
Loula Rytikova, *George Mason University*
Fary Sami, *Harford Community College*
Vicki Schell, *Pensacola State College*
Carol Schoen, *University of Wisconsin–Eau Claire*
Pali Sen, *University of North Florida*
Rosa Seyfried, *Harrisburg Area Community College*
Larry Shrewsbury, *Southern Oregon University*
Abdallah Shuaibi, *Truman College*
Rick Silvey, *University of Saint Mary*
Russell Simmons, *Brookhaven College*
Peggy Slavik, *University of Saint Mary*
Karen Smith, *University of West Georgia*
Pam Stogsdill, *Bossier Parish Community College*
Susan Surina, *George Mason University*
Victor Swaim, *Southeastern Louisiana University*
Scott Sykes, *University of West Georgia*

Van Tran, *San Francisco State University*
John Trimboli, *Middle Georgia State University*
Barbara Tucker, *Tarrant County College South East*
Steven Forbes Tuckey, *Jackson Community College*
Christopher Turner, *Pensacola State College*
Anke Van Zuylen, *College of William and Mary*
Dave Vinson, *Pellissippi State Community College*
Joseph Walker, *Georgia State University*
James Wan, *Long Beach City College*
Xiaohong Wang, *Central Michigan University*
Jason Willis, *Gardner-Webb University*
Fuzhen Zhang, *Nova Southeastern University*
Yichuan Zhao, *Georgia State University*
Deborah Ziegler, *Hannibal LaGrange University*
Bashar Zogheib, *Nova Southeastern University*
Stephanie Zwyghuizen, *Jamestown Community College*

McGraw Hill ALEKS®

Personalized for your students,
flexible for your approach.

Every student has different needs and enters your course with varied levels of preparation. ALEKS® pinpoints what students already know, what they don't and, most importantly, what they're ready to learn next. ALEKS® provides flexibility and control when shaping your course so you can reach every student.

ALEKS® Keeps Students Focused with a Personalized Learning Path

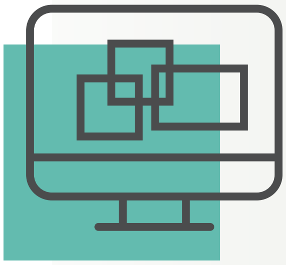
ALEKS® creates an optimized path with an ongoing cycle of learning and assessment, celebrating students' small wins along the way with positive real-time feedback. Rooted in research and analytics, ALEKS® improves student outcomes by fostering better preparation, increased motivation and knowledge retention.

Your Course, Your Way

In a single platform, ALEKS® provides a balance of personalized, adaptive practice and non-adaptive assignments for application and assessment. This supports students individually while allowing you to create assignments that align to your course goals.

By differentiating instruction for students, ALEKS® unlocks the power of instructional time, group work, and applications and prepares students to make real-world connections.

//CODiE//
2019 SIIA CODiE WINNER



Flexible Implementation

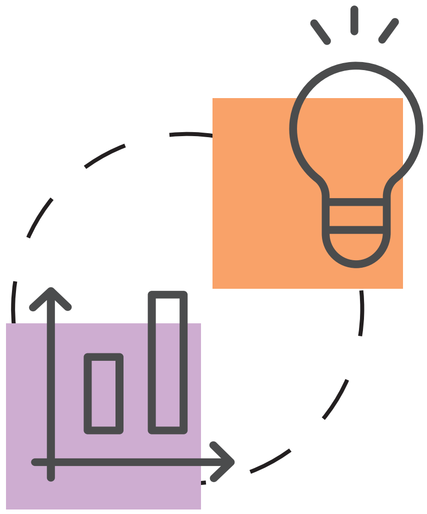
ALEKS® enables you to structure your course regardless of your instruction style and format. From a traditional classroom, to various corequisite models, to an online prep course before the start of the term, ALEKS® can supplement your instruction or play a lead role in delivering the content.

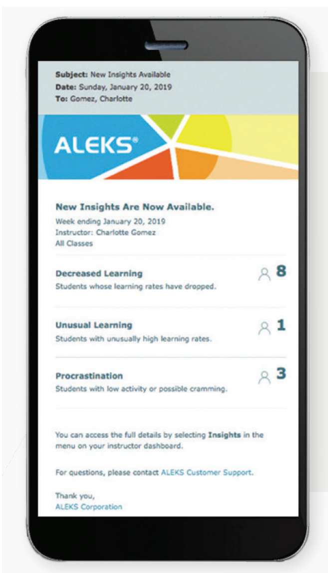
Teaching online? Respondus integration with ALEKS® supports secure testing with LockDown Browser and Respondus Monitor for Tests, Quizzes, and Scheduled Knowledge Checks.

Outcomes & Efficacy

Our commitment to improve student outcomes services a wide variety of implementation models and best practices, from lecture-based to labs and corequisites to summer prep courses. Our case studies illustrate our commitment to help you reach your course goals and our research demonstrates our drive to support all students, regardless of their math background and preparation level.

*visit **bit.ly/outcomesandefficacy** to review empirical data from ALEKS® users around the country





Turn Data Into Actionable Insights

ALEKS® Reports are designed to inform your instruction and create more meaningful interactions with your students when they need it the most. ALEKS® Insights alert you when students might be at risk of falling behind so that you can take immediate action. Insights summarize students exhibiting at least one of four negative behaviors that may require intervention including Failed Topics, Decreased Learning, Unusual Learning and Procrastination & Cramming.

DIGITAL EDGE50 AWARDS 2019

Winner of 2019 Digital Edge 50 Award for Data Analytics!

bit.ly/ALEKS_MHE

Supplements

Multimedia Supplements

This text is available with the adaptive online learning platform ALEKS. Rooted in 20 years of research and analytics, ALEKS identifies what students know, what they don't know, and what they are ready to learn, then personalizes a learning path tailored to each student. This personalization keeps students focused on what they need to learn and prepares them to make real-world connections. ALEKS also delivers textbook-aligned exercises, videos, and the many author-created digital supplements that accompany this text.

With **McGraw Hill Create**™, you can easily rearrange chapters, combine material from other content sources, and quickly upload content you have written like your course syllabus or teaching notes. Find the content you need in Create by searching through thousands of leading McGraw Hill textbooks. Arrange your book to fit your teaching style. Create even allows you to personalize your book's appearance by selecting the cover and adding your name, school, and course information. Assemble a Create book, and you'll receive a complimentary print review copy in 3–5 business days or a complimentary electronic review copy (eComp) via email in minutes. Go to www.mcgrawhillcreate.com today and experience how McGraw Hill Create™ empowers you to teach *your* students *your* way.

Computerized Test Bank Online (instructors only)

This computerized test bank, available online to adopting instructors, utilizes TestGen® cross-platform test generation software to quickly and easily create customized exams. Using hundreds of test items taken directly from the text, TestGen allows rapid test creation and flexibility for instructors to create their own questions from scratch with the ability to randomize number values. Powerful search and sort functions help quickly locate questions and arrange them in any order, and built-in mathematical templates let instructors insert stylized text, symbols, graphics, and equations directly into questions without need for a separate equation editor.

Videos

Author-produced lecture videos introduce concepts, definitions, formulas, and problem-solving procedures to help students better comprehend the topic at hand. Exercise videos illustrate the authors working through selected exercises, following the solution methodology employed in the text. These videos are closed-captioned for the hearing-impaired and meet the Americans with Disabilities Act Standards for Accessible Design.

SPSS Student Version for Windows

A student version of SPSS statistical software is available with copies of this text. Consult your McGraw Hill representative for details.

Guided Student Notes

Guided notes provide instructors with the framework of day-by-day class activities for each section in the book. Each lecture guide can help instructors make more efficient use of class time and can help keep students focused on active learning. Students who use the lecture guides have the framework of well-organized notes that can be completed with the instructor in class.

Data Sets

Data sets from selected exercises have been pre-populated into MINITAB, TI-Graph Link, Excel, SPSS, and comma-delimited ASCII formats for student and instructor use. These files are available on the text's website.

MINITAB 17 Manual

With guidance from the authors, this manual includes material from the book to provide seamless use from one to the other, providing additional practice in applying the chapter concepts while using the MINITAB program.

TI-84 Plus Graphing Calculator Manual

This friendly, author-influenced manual teaches students to learn about statistics and solve problems by using this calculator while following each text chapter.

Excel Manual

This workbook, specially designed to accompany the text by the authors, provides additional practice in applying the chapter concepts while using Excel.

Print Supplements

Annotated Instructor's Edition (instructors only)

The Annotated Instructor's Edition contains answers to all exercises. The answers to most questions are printed in blue next to each problem. Answers not appearing on the page can be found in the Answer Appendix at the end of the book.

Statistics Corequisite Workbook

This workbook, written by co-author Barry Monk, is designed to provide corequisite remediation of the necessary skills for an introductory statistics course. The included topics are largely independent of one another and may be used in any order that works best for the instructor. The workbook is available online or can be ordered in print format through Create.

Index of Applications

Elderly individuals applications

Electronics/computer applications

Entertainment applications. *See also* Gambling applications

Environmental applications

Exercise applications. *See* Sports/exercise/fitness applications

Farming applications. *See* Agricultural/gardening/farming applications

Food applications

Gambling applications

Gardening applications. *See* Agricultural/gardening/farming applications

Genetics/gender applications

Health/healthcare applications. *See also* Medical/medical research/alternative medicine applications

Tatiana Grozetskaya/Shutterstock

chapter 1

Basic Ideas

Introduction

How does air pollution affect your health? Over the past several decades, scientists have become increasingly convinced that air pollution is a serious health hazard. The World Health Organization has estimated that air pollution causes 2.4 million deaths each year. The health effects of air pollution have been investigated by measuring air pollution levels and rates of disease, then using statistical methods to determine whether higher levels of pollution lead to higher rates of disease.

Many air pollution studies have been conducted in the United States. For example, the town of Libby, Montana, was the focus of a recent study of the effect of particulate matter—air pollution that consists of microscopic particles—on the respiratory health of children. As part of this study, parents were asked to fill out a questionnaire about their children's respiratory symptoms. It turned out that children exposed to higher levels of particulate pollution were more likely to exhibit symptoms of wheezing, as shown in the following table.

Level of Exposure	Percentage with Symptoms
High	8.89%
Low	4.56%

The rate of symptoms was almost twice as high among those exposed to higher levels of pollution. At first, it might seem easy to conclude that higher levels of pollution cause symptoms of wheezing. However, drawing accurate conclusions from information like this is rarely that simple. The case study at the end of this chapter will present more complete information and will show that additional factors must be considered.

SECTION 1.1 Sampling

Objectives

1. Describe the investigative process of statistics
2. Construct a simple random sample
3. Determine when samples of convenience are acceptable
4. Describe stratified sampling, cluster sampling, systematic sampling, and voluntary response sampling
5. Distinguish between statistics and parameters

Objective 1 Describe the investigative process of statistics

In the months leading up to an election, polls often tell us the percentages of voters that prefer each of the candidates. How do pollsters obtain this information? The ideal poll would be one in which every registered voter were asked his or her opinion. Of course, it is impossible to conduct such an ideal poll, because it is impossible to contact every voter. Instead, pollsters contact a relatively small number of voters, usually no more than a couple of thousand, and use the information from these voters to predict the preferences of the entire group of voters.

The process of polling requires two major steps. First, the voters to be polled must be selected and interviewed. In this way the pollsters collect information. In the second step, the pollsters analyze the information to make predictions about the upcoming election. Both the collection and the analysis of the information must be done properly for the results to be reliable. The field of statistics provides appropriate methods for the collection, description, and analysis of information.

DEFINITION

Statistics is the study of procedures for collecting, describing, and drawing conclusions from information.

Polling follows a process that is typical in statistics. We formulate questions (Which candidate is most likely to win?), then collect and analyze data to address the questions. In general, statistics is an investigative process that involves the following four steps:

- Formulate questions.
- Collect data needed to answer the questions.
- Describe the data.
- Draw conclusions, using appropriate methods.

Explain It Again

Why do we draw samples? It's usually impossible to examine every member of a large population. So we select a group of a manageable size to examine. This group is called a sample.

In this section, we will focus on the second step in the process—the collection of data. The polling problem is typical of a data collection problem. We want some information about a large group of individuals, but we are able to collect information on only a small part of that group. In statistical terminology, the large group is called a *population*, and the part of the group on which we collect information is called a *sample*.

DEFINITION

- A **population** is the entire collection of individuals about which information is sought.
- A **sample** is a subset of a population, containing the individuals that are actually observed.

Ideally, we would like our sample to represent the population as closely as possible. For example, in a political poll, we would like the proportions of voters preferring each of the candidates to be the same in the sample as in the population. Unfortunately, there are no methods that can guarantee that a sample will represent the population well. The best we can do is to use a method that makes it very likely that the sample will be similar to the population. The best sampling methods all involve some kind of random selection. The most basic, and in many cases the best, sampling method is the method of **simple random sampling**.

Objective 2 Construct a simple random sample

Simple Random Sampling

To understand the nature of a simple random sample, think of a lottery. Imagine that 10,000 lottery tickets have been sold, and that 5 winners are to be chosen. What is the fairest way to choose the winners? The fairest way is to put the 10,000 tickets in a drum, mix them thoroughly, then reach in and draw 5 tickets out one by one. These 5 winning tickets are a simple random sample from the population of 10,000 lottery tickets. Each ticket is equally likely to be one of the 5 tickets drawn. More importantly, each collection of 5 tickets that can be formed from the 10,000 is equally likely to comprise the group of 5 that is drawn.

DEFINITION

A **simple random sample** of size n is a sample chosen by a method in which each collection of n population items is equally likely to make up the sample, just as in a lottery.

Since a simple random sample is analogous to a lottery, it can often be drawn by the same method now used in many lotteries: with a computer random number generator. Suppose there are N items in the population. We number the items 1 through N. Then we generate a list of random integers between 1 and N and choose the corresponding population items to comprise the simple random sample.

EXAMPLE 1.1

Choosing a simple random sample

There are 300 employees in a certain company. The Human Resources department wants to draw a simple random sample of 20 employees to fill out a questionnaire about their attitudes toward their jobs. Describe how technology can be used to draw this sample.

Solution

Step 1: Make a list of all 300 employees, and number them from 1 to 300.

Step 2: Use a random number generator on a computer or a calculator to generate 20 random numbers between 1 and 300. The employees who correspond to these numbers comprise the sample.

EXAMPLE 1.2

Determining whether a sample is a simple random sample

A physical education professor wants to study the physical fitness levels of students at her university. There are 20,000 students enrolled at the university, and she wants to draw a sample of size 100 to take a physical fitness test. She obtains a list of all 20,000 students, numbered from 1 to 20,000. She uses a computer random number generator to generate 100 random integers between 1 and 20,000, then invites the 100 students corresponding to those numbers to participate in the study. Is this a simple random sample?

Solution

Yes, this is a simple random sample because any group of 100 students would have been equally likely to have been chosen.

EXAMPLE 1.3

Determining whether a sample is a simple random sample

The professor in Example 1.2 now wants to draw a sample of 50 students to fill out a questionnaire about which sports they play. The professor's 10:00 A.M. class has 50 students. She uses the first 20 minutes of class to have the students fill out the questionnaire. Is this a simple random sample?

Solution

No. A simple random sample is like a lottery, in which each student in the population has an equal chance to be part of the sample. In this case, only the students in a particular class had a chance to be in the sample.

EXAMPLE 1.4 In a simple random sample, all samples are equally likely

To play the Colorado Lottery Lotto game, you must select six numbers from 1 to 42. Then lottery officials draw a simple random sample of six numbers from 1 to 42. If your six numbers match the ones in the simple random sample, you win the jackpot. Sally plays the lottery and chooses the numbers 1, 2, 3, 4, 5, 6. Her friend George says that this isn't a good choice, since it is very unlikely that a random sample will turn up the first six numbers. Is he right?

Solution

No. It is true that the combination 1, 2, 3, 4, 5, 6 is unlikely, but every other combination is equally unlikely. In a simple random sample of size 6, every collection of six numbers is equally likely (or equally unlikely) to come up. So Sally has the same chance as anyone to win the jackpot.

EXAMPLE 1.5 Using technology to draw a simple random sample

Use technology to draw a simple random sample of five employees from the following list.

1. Dan Aaron	11. Johnny Gaines	21. Jorge Ibarra	31. Edward Shingleton
2. Annie Bienh	12. Carlos Garcia	22. Maurice Jones	32. Michael Speciale
3. Oscar Bolivar	13. Julio Gonzalez	23. Jared Kerns	33. Andrew Steele
4. Dominique Bonnaud	14. Jacqueline Gordon	24. Kevin King	34. Neil Swain
5. Paul Campbell	15. James Graves	25. Frank Lipka	35. Sherry Thomas
6. Jeffrey Carnahan	16. Ronald Harrison	26. Carl Luther	36. Shequiea Thompson
7. Joel Chae	17. Andrew Huang	27. Laverne Mitchell	37. Barbara Tilford
8. Dustin Chen	18. Anthony Hunter	28. Zachary Quesada	38. Jermaine Tryon
9. Steven Coleman	19. Jonathan Jackson	29. Donnell Romaine	39. Lizbet Valdez
10. Richard Davis	20. Bruce Johnson	30. Gary Sanders	40. Katelyn Yu

Solution

We will use the TI-84 Plus graphing calculator. The step-by-step procedure is presented in the Using Technology section on page 9. We begin by choosing a **seed**, which is a number that the calculator uses to get the random number generator started. Display (a) shows the seed being set to 21. (The seed can be chosen in almost any way; this number was chosen by looking at the seconds display on a digital watch.) Display (b) presents the five numbers in the sample.

```
21→rand
```

(a)

```
randIntNoRep(1,40,5)
              {27 39 30 35 17}
```

(b)

The simple random sample consists of the employees with numbers 27, 39, 30, 35, and 17. These are Laverne Mitchell, Lizbet Valdez, Gary Sanders, Sherry Thomas, and Andrew Huang.

CAUTION

If you use a different type of calculator, a different statistical package, or a different seed, you will get a different random sample. This is perfectly all right. So long as the sample is drawn by using a correct procedure, it is a valid random sample.

Check Your Understanding

1. A pollster wants to estimate the proportion of voters in a certain town who are Democrats. He goes to a large shopping mall and approaches people to ask whether they are Democrats. Is this a simple random sample? Explain.

2. A telephone company wants to estimate the proportion of customers who are satisfied with their service. They use a computer to generate a list of random phone numbers and call those people to ask them whether they are satisfied. Is this a simple random sample? Explain.

Answers are on page 12.

Objective 3 Determine when samples of convenience are acceptable

Samples of Convenience

In some cases, it is difficult or impossible to draw a sample in a truly random way. In these cases, the best one can do is to sample items by some convenient method. A sample obtained in such a way is called a *sample of convenience*.

DEFINITION

A **sample of convenience** is a sample that is not drawn by a well-defined random method.

EXAMPLE 1.6

Creatas Images/Jupiterimages

Drawing a sample of convenience

A construction engineer has just received a shipment of 1000 concrete blocks, each weighing approximately 50 pounds. The blocks have been delivered in a large pile. The engineer wishes to investigate the crushing strength of the blocks by measuring the strengths in a sample of 10 blocks. Explain why it might be difficult to draw a simple random sample of blocks. Describe how the engineer might draw a sample of convenience.

Solution

To draw a simple random sample would require removing blocks from the center and bottom of the pile, which might be quite difficult. One way to draw a sample of convenience would be to simply take 10 blocks off the top of the pile.

Problems with samples of convenience

The big problem with samples of convenience is that they may differ systematically in some way from the population. For this reason, samples of convenience should not be used, except in some situations where it is not feasible to draw a random sample. When it is necessary to draw a sample of convenience, it is important to think carefully about all the ways in which the sample might differ systematically from the population. If it is reasonable to believe that no important systematic difference exists, then it may be acceptable to treat the sample of convenience as if it were a simple random sample. With regard to the concrete blocks, if the engineer is confident that the blocks on the top of the pile do not differ systematically in any important way from the rest, then he can treat the sample of convenience as a simple random sample. If, however, it is possible that blocks in different parts of the pile may have been made from different batches of mix, or may have different curing times or temperatures, a sample of convenience could give misleading results.

CAUTION

Don't use a sample of convenience when it is possible to draw a simple random sample.

SUMMARY

- A sample of convenience may be acceptable when it is reasonable to believe that there is no systematic difference between the sample and the population.
- A sample of convenience is not acceptable when it is possible that there is a systematic difference between the sample and the population.

Objective 4 Describe stratified sampling, cluster sampling, systematic sampling, and voluntary response sampling

Some Other Sampling Methods

Stratified sampling

In **stratified sampling**, the population is divided into groups, called **strata**, where the members of each stratum are similar in some way. Then a simple random sample is drawn from each stratum. Stratified sampling is useful when the strata differ from one another, but the individuals within a stratum tend to be alike.

EXAMPLE 1.7

Drawing a stratified sample

A company has 1000 employees, of whom 800 are full-time and 200 are part-time. The company wants to survey 50 employees about their opinions regarding benefits. Attitudes toward benefits may differ considerably between full-time and part-time employees. Why might it be a good idea to draw a stratified sample? Describe how one might be drawn.

Solution

If a simple random sample is drawn from the entire population of 1000 employees, it is possible that the sample will contain only a few part-time employees, and their attitudes will not be well represented. For this reason, it might be advantageous to draw a stratified sample. To draw a stratified sample, one would use two strata. One stratum would consist of the full-time employees, and the other would consist of the part-time employees. A simple random sample would be drawn from the full-time employees, and another simple random sample would be drawn from the part-time employees. This method guarantees that both full-time and part-time employees will be well represented.

Explain It Again

Example of a cluster sample: Imagine drawing a simple random sample of households and interviewing every member of each household. This would be a cluster sample, with the households as the clusters.

Cluster sampling

In **cluster sampling**, items are drawn from the population in groups, or clusters. Cluster sampling is useful when the population is too large and spread out for simple random sampling to be feasible. Cluster sampling is used extensively by U.S. government agencies in sampling the U.S. population to measure sociological factors such as income and unemployment.

EXAMPLE 1.8

Drawing a cluster sample

To estimate the unemployment rate in a county, a government agency draws a simple random sample of households in the county. Someone visits each household and asks how many adults live in the household and how many of them are unemployed. What are the clusters? Why is this a cluster sample?

Solution

The clusters are the groups of adults in each of the households in the county. This is a cluster sample because a simple random sample of clusters is selected, and every individual in each selected cluster is part of the sample.

Explain It Again

The difference between cluster sampling and stratified sampling: In both cluster sampling and stratified sampling, the population is divided into groups. In stratified sampling, a simple random sample is chosen from each group. In cluster sampling, a random sample of groups is chosen, and every member of the chosen groups is sampled.

Systematic sampling

Imagine walking alongside a line of people and choosing every third one. That would produce a **systematic sample**. In a systematic sample, the population items are ordered. It is decided how frequently to sample items; for example, one could sample every third item, or every fifth item, or every hundredth item. Let k represent the sampling frequency. To begin the sampling, choose a starting place at random. Select the item in the starting place, along with every kth item after that.

Systematic sampling is sometimes used to sample products as they come off an assembly line, in order to check that they meet quality standards.

EXAMPLE 1.9

Describe a systematic sample

Automobiles are coming off an assembly line. It is decided to draw a systematic sample for a detailed check of the steering system. The starting point will be the third car, then every fifth car after that will be sampled. Which cars will be sampled?

Digital Vision/Punchstock

Solution

We start with the third car, then count by fives to determine which cars will be sampled. The sample will consist of cars numbered 3, 8, 13, 18, and so on.

Voluntary response sampling

Voluntary response samples are often used by the media to try to engage the audience. For example, a news commentator will invite people to tweet an opinion, or a radio announcer will invite people to call the station to say what they think. How reliable are voluntary response samples? To put it simply, *voluntary response samples are never reliable*. People who go to the trouble to volunteer an opinion tend to have stronger opinions than is typical of the population. In addition, people with negative opinions are often more likely to volunteer their responses than those with positive opinions.

Figures 1.1–1.4 illustrate several valid methods of sampling.

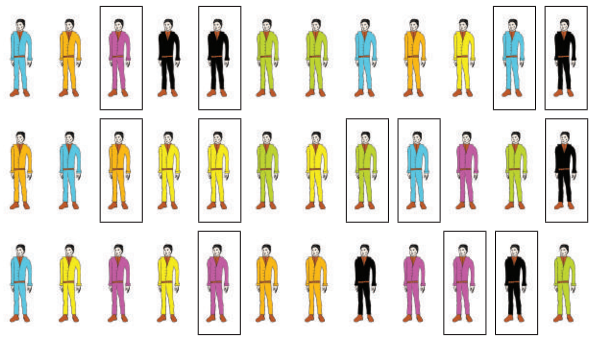


Figure 1.1 Simple random sampling

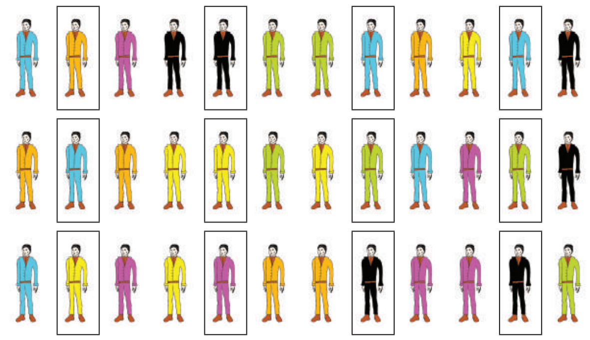


Figure 1.2 Systematic sampling

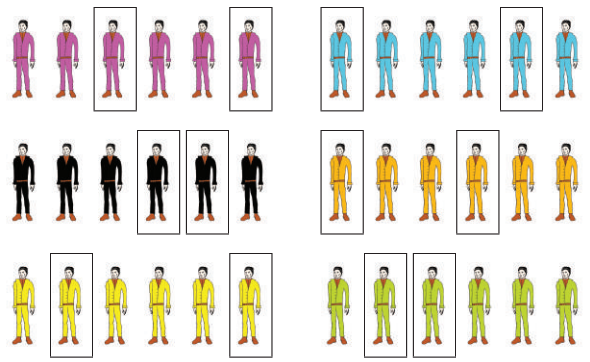


Figure 1.3 Stratified sampling

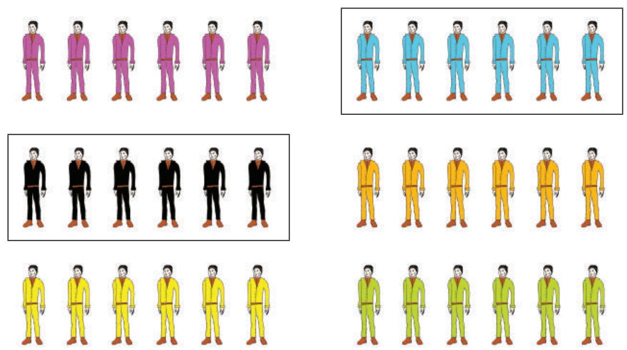


Figure 1.4 Cluster sampling

Check Your Understanding

3. A radio talk show host invites listeners to send an email to express their opinions on an upcoming election. More than 10,000 emails are received. What kind of sample is this?

4. Every 10 years, the U.S. Census Bureau attempts to count every person living in the United States. To check the accuracy of their count in a certain city, they draw a sample of census districts (roughly equivalent to a city block) and recount everyone in the sampled districts. What kind of sample is formed by the people who are recounted?

5. A public health researcher is designing a study of the effect of diet on heart disease. The researcher knows that the diets of men and women tend to differ and that men are more susceptible to heart disease. To be sure that both men and women are well represented, the study comprises a simple random sample of 100 men and another simple random sample of 100 women. What kind of sample do these 200 people represent?
6. A college basketball team held a promotion at one of its games in which every twentieth person who entered the arena won a free basketball. What kind of sample do the winners represent?

Answers are on page 12.

Simple random sampling is the most basic method

Simple random sampling is not the only valid method of random sampling. But it is the most basic, and we will focus most of our attention on this method. From now on, unless otherwise stated, the terms *sample* and *random sample* will be taken to mean *simple random sample*.

Objective 5 Distinguish between statistics and parameters

Statistics and Parameters

We often use numbers to describe, or summarize, a sample or a population. For example, suppose that a pollster draws a sample of 500 likely voters in an upcoming election, and 68% of them say that the state of the economy is the most important issue for them. The quantity "68%" describes the sample. A number that describes a sample is called a *statistic*.

DEFINITION

A **statistic** is a number that describes a sample.

Explain It Again

Statistic and parameter: An easy way to remember these terms is that "statistic" and "sample" both begin with "s," and "parameter" and "population" both begin with "p."

Now imagine that the election takes place and that one of the items on the ballot is a proposition to raise the sales tax to pay for the development of a new park downtown. Let's say that 53% of the voters vote in favor of the proposition. The quantity "53%" describes the population of voters who voted in the election. A number that describes a population is called a *parameter*.

DEFINITION

A **parameter** is a number that describes a population.

EXAMPLE 1.10

Distinguishing between a statistic and a parameter

Which of the following is a statistic and which is a parameter?

a. 57% of the teachers at Central High School are female.
b. In a sample of 100 surgery patients who were given a new pain reliever, 78% of them reported significant pain relief.

Solution

a. The number 57% is a parameter, because it describes the entire population of teachers in the school.
b. The number 78% is a statistic, because it describes a sample.

USING TECHNOLOGY

We use Example 1.5 to illustrate the technology steps.

TI-84 PLUS

Drawing a simple random sample

Step 1. Enter any nonzero number on the HOME screen as the seed.

Step 2. Press **STO >**.

Step 3. Press **MATH**, select **PRB**, then **1: rand**, and then press **ENTER**. This enters the seed into the calculator memory. See Figure A, which uses the number 21 as the seed.

Step 4. Press **MATH**, select **PRB**, then **5: randIntNoRep**. Then enter **1, N, n**, where **N** is the population size and **n** is the desired sample size. In Example 1.5, we use **N** = 40 and **n** = 5 (Figure B).

Step 5. Press **ENTER**. The five values in the random sample for Example 1.5 are **27, 39, 30, 35, 17** (Figure C).

Figure A

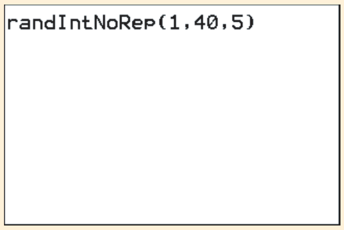


Figure B

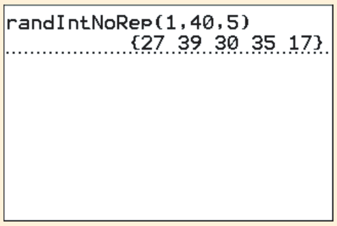


Figure C

EXCEL

Drawing a simple random sample

Step 1. In **Column A**, enter the values **1** through the population size **N**. For Example 1.5, **N** = 40.

Step 2. In **Column B**, next to each value in **Column A**, enter the command **=RAND( )**. This results (Figure D) in a randomly generated number between 0 and 1 in each cell in **Column B**.

Step 3. Select all values in **Columns A** and **B** and then click on the **Data** menu and select **Sort**.

Step 4. In the **Sort by** field, enter **Column B** and select **Smallest to Largest** in the **Order** field. Press **OK**. **Column A** now contains the random sample. Our random sample begins with **17, 12, 28, 20, 6, ...** (Figure E).

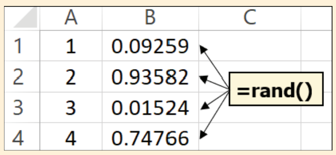


Figure D

	A	B
1	17	0.9198
2	12	0.83333
3	28	0.1054
4	20	0.8635
5	6	0.82038
6	11	0.51474
7	36	0.05115
8	19	0.75697
9	39	0.82656
10	26	0.31556
11	8	0.11958
12	30	0.32979
13	35	0.59734

Figure E

MINITAB

Drawing a simple random sample

Step 1. Click **Calc**, then **Random Data**, then **Integer...**

Step 2. In the **Number of rows of data to generate** field, enter twice the desired sample size. For example, if the desired sample size is 10, enter 20. The reason for this is that some sample items may be repeated, and these will need to be deleted.

Step 3. In the **Store in column(s)** field, enter **C1**.

Step 4. Enter **1** for the **Minimum value** and the population size **N** for the **Maximum value**. We use **Maximum value** = 40 for Example 1.5. Click **OK**.

Step 5. **Column C1** of the worksheet will contain a list of randomly selected numbers between **1** and **N**. If any number appears more than once in **Column C1**, delete the replicates so that the number appears only once. For Example 1.5, our random sample begins with **16, 14, 30, 28, 17, ...** (Figure F).

↓	C1
1	16
2	14
3	30
4	28
5	17
6	13
7	4
8	8
9	6
10	15
11	35
12	5

Figure F

SECTION 1.1 Exercises

Exercises 1–6 are the Check Your Understanding exercises located within the section.

Understanding the Concepts

In Exercises 7–12, fill in each blank with the appropriate word or phrase.

7. The entire collection of individuals about which information is sought is called a ________.

8. A ________ is a subset of a population.

9. A ________ is a type of sample that is analogous to a lottery.

10. A sample that is not drawn by a well-defined random method is called a ________.

11. A ________ sample is one in which the population is divided into groups and a random sample of groups is drawn.

12. A ________ sample is one in which the population is divided into groups and a random sample is drawn from each group.

In Exercises 13–16, determine whether the statement is true or false. If the statement is false, rewrite it as a true statement.

13. A sample of convenience is never acceptable.

14. In a cluster sample, the population is divided into groups, and a random sample from each group is drawn.

15. Both stratified sampling and cluster sampling divide the population into groups.

16. One reason that voluntary response sampling is unreliable is that people with stronger views tend to express them more readily.

Practicing the Skills

In Exercises 17–22, determine whether the number described is a statistic or a parameter.

17. In a recent poll, 57% of the respondents supported a school bond issue.

18. The average age of the employees in a certain company is 35 years.

19. Of the students enrolled in a certain college, 80% are full-time.

20. In a survey of 500 high school students, 60% of them said that they intended to go to college.

21. The U.S. Census reports that 95% of California residents live in urban areas.

22. In a survey of parents with children under the age of six, 90% said that their child had been seen by a pediatrician within the past year.

Exercises 23–26 refer to the population of animals in the following table. The population is divided into four groups: mammals, birds, reptiles, and fish.

Mammals		Birds	
1. Aardvark	6. Lion	11. Flamingo	16. Hawk
2. Buffalo	7. Zebra	12. Swan	17. Owl
3. Elephant	8. Pig	13. Sparrow	18. Chicken
4. Squirrel	9. Dog	14. Parrot	19. Duck
5. Rabbit	10. Horse	15. Pelican	20. Turkey
Reptiles		**Fish**	
21. Gecko	26. Python	31. Catfish	36. Shark
22. Iguana	27. Turtle	32. Tuna	37. Trout
23. Chameleon	28. Tortoise	33. Cod	38. Perch
24. Rattlesnake	29. Alligator	34. Salmon	39. Guppy
25. Boa constrictor	30. Crocodile	35. Goldfish	40. Minnow

23. **Simple random sample:** Draw a simple random sample of eight animals from the list of 40 animals in the table.
24. **Another sample:** Draw a sample of eight animals by drawing a simple random sample of two animals from each group. What kind of sample is this?
25. **Another sample:** Draw a simple random sample of two groups of animals from the four groups, and construct a sample of 20 animals by including all the animals in the sampled groups. What kind of sample is this?
26. **Another sample:** Choose a random number between 1 and 5. Include the animal with that number in your sample, along with every fifth animal thereafter, to construct a sample of eight animals. What kind of sample is this?

In Exercises 27–42, identify the kind of sample that is described.

27. **Parking on campus:** A college faculty consists of 400 men and 250 women. The college administration wants to draw a sample of 65 faculty members to ask their opinion about a new parking fee. They draw a simple random sample of 40 men and another simple random sample of 25 women.
28. **Cruising the mall:** A pollster walks around a busy shopping mall and approaches people passing by to ask them how often they shop at the mall.
29. **What's on TV?** A pollster obtains a list of all the residential addresses in a certain town and uses a computer random number generator to choose 150 of them. The pollster visits each of the 150 households and interviews all the adults in each household about their television viewing habits.
30. **Don't drink and drive:** Police at a sobriety checkpoint pull over every fifth car to determine whether the driver is sober.
31. **Tell us your opinion:** A television newscaster invites viewers to tweet their opinions on a proposed bill on immigration policy. More than 50,000 people express their opinions in this way.
32. **Reading program:** The superintendent of a large school district wants to test the effectiveness of a new program designed to improve reading skills among elementary school children. There are 30 elementary schools in the district. The superintendent chooses a simple random sample of five schools and institutes the new reading program in those schools. A total of 4700 children attend these five schools.

Somos Images/Alamy

33. **Customer survey:** All the customers who entered a store on a particular day were given a survey to fill out concerning their opinions of the service at the store.
34. **Raffle:** Five hundred people attend a charity event, and each buys a raffle ticket. The 500 ticket stubs are put in a drum and thoroughly mixed, and 10 of them are drawn. The 10 people whose tickets are drawn win a prize.
35. **Hospital survey:** The director of a hospital pharmacy chooses at random 100 people age 60 or older from each of three surrounding counties to ask their opinions of a new prescription drug program.
36. **Bus schedule:** Officials at a metropolitan transit authority want to get input from people who use a certain bus route about a possible change in the schedule. They randomly select 5 buses during a certain week and poll all riders on those buses about the change.
37. **How much did you spend?** A retailer samples 25 receipts from the past week by numbering all the receipts, generating 25 random numbers, and sampling the receipts that correspond to these numbers.
38. **Phone features:** A phone company wants to draw a sample of 600 customers to gather opinions about potential new features on upcoming phone models. The company draws a random sample of 200 from customers with iPhones, a random sample of 100 from customers with LG phones, a random sample of 100 from customers with Samsung phones, and a random sample of 200 from customers with other phones.
39. **Computer network:** Every third day, a computer network administrator analyzes the company's network logs to check for signs of computer viruses.
40. **Apps:** An app produces a message requesting customers to click on a link to rate the app.
41. **Survey:** A nutritionist randomly chooses 100 people who have been using a certain weight loss program and asks them how much weight they have lost.
42. **Good service:** Receipts at a department store indicate a website where customers can take a survey indicating their level of satisfaction with the service they received.

Working with the Concepts

43. **You're giving me a headache:** A pharmaceutical company wants to test a new drug that is designed to provide superior relief from headaches. They want to select a sample of headache sufferers to try the drug. Do you think that it is feasible to draw a simple random sample of headache sufferers, or will it be necessary to use a sample of convenience? Explain your reasoning.

44. **Pay more for recreation?** The director of the recreation center at a large university wants to sample 100 students to ask them whether they would support an increase in their recreation fees in order to expand the hours that the center is open. Do you think that it is feasible to draw a simple random sample of students, or will it be necessary to use a sample of convenience? Explain your reasoning.

45. **Voter preferences:** A pollster wants to sample 500 voters in a town to ask them who they plan to vote for in an upcoming election. Describe a sampling method that would be appropriate in this situation. Explain your reasoning.

46. **Quality control:** Products come off an assembly line at the rate of several hundred per hour. It is desired to sample 10% of them to check whether they meet quality standards. Describe a sampling method that would be appropriate in this situation. Explain your reasoning.

47. **On-site day care:** A large company wants to sample 200 employees to ask their opinions about providing a day care center for the employees' children. They want to be sure to sample equal numbers of men and women. Describe a sampling method that would be appropriate in this situation. Explain your reasoning.

48. **The tax man cometh:** The Internal Revenue Service wants to sample 1000 tax returns that were submitted last year to determine the percentage of returns that had a refund. Describe a sampling method that would be appropriate in this situation. Explain your reasoning.

Extending the Concepts

49. **Draw a sample:** Imagine that you are asked to determine students' opinions at your school about a potential change in library hours. Describe how you could go about getting a sample of each of the following types: simple random sample, sample of convenience, voluntary response sample, stratified sample, cluster sample, systematic sample.

50. **A systematic sample is a cluster sample:** Explain how a systematic sample is actually a type of cluster sample.

Answers to Check Your Understanding Exercises for Section 1.1

1. No; this sample consists only of people in the town who visit the mall.
2. Yes; every group of n customers, where n is the sample size, is equally likely to be chosen.
3. Voluntary response sample
4. Cluster sample
5. Stratified sample
6. Systematic sample

SECTION 1.2 Types of Data

Objectives

1. Understand the structure of a typical data set
2. Distinguish between qualitative and quantitative variables
3. Distinguish between ordinal and nominal variables
4. Distinguish between discrete and continuous variables
5. Distinguish between ratio and interval levels of measurement

Objective 1 Understand the structure of a typical data set

Data Sets

In Section 1.1, we described various methods of collecting information by sampling. Once the information has been collected, the collection is called a **data set**. A simple example of a data set is presented in Table 1.1, which shows the major, final exam score, and grade for several students in a certain statistics class.


Comstock/Stockbyte/Getty Images

Table 1.1 Major, Final Exam Score, and Grade for Several Students

Student	Major	Exam Score	Grade
1	Psychology	92	A
2	Business	75	B
3	Communications	82	B
4	Psychology	72	C
5	Art	85	B

Table 1.1 illustrates some basic features that are found in most data sets. Information is collected on **individuals**. In this example, the individuals are students. In many cases, individuals are people; in other cases, they can be animals, plants, or things. The characteristics of the individuals about which we collect information are called **variables**.

In this example, the variables are major, exam score, and grade. Finally, the values of the variables that we obtain are the **data**. So, for example, the data for individual #1 are Major = Psychology, Exam score = 92, and Grade = A.

Check Your Understanding

1. A pollster asks a group of six voters about their political affiliation (Republican, Democrat, or Independent), their age, and whether they voted in the last election. The results are shown in the following table.

Voter	Political Affiliation	Age	Voted in Last Election?
1	Republican	34	Yes
2	Democrat	56	Yes
3	Democrat	21	No
4	Independent	28	Yes
5	Republican	61	No
6	Independent	46	Yes

a. How many individuals are there?
b. Identify the variables.
c. What are the data for individual #3?

Answers are on page 19.

Objective 2 Distinguish between qualitative and quantitative variables

Qualitative and Quantitative Variables

Variables can be divided into two types: qualitative and quantitative. **Qualitative variables**, also called **categorical variables**, classify individuals into categories. For example, college major and gender are qualitative variables. **Quantitative variables** are numerical and tell how much or how many of something there is. Height and score on an exam are examples of quantitative variables.

Explain It Again

Another way to distinguish qualitative from quantitative variables: Quantitative variables are counts or measurements, whereas qualitative variables are descriptions.

SUMMARY

- Qualitative variables classify individuals into categories.
- Quantitative variables tell how much or how many of something there is.

EXAMPLE 1.11

Distinguishing between qualitative and quantitative variables

Which of the following variables are qualitative and which are quantitative?

a. A person's age
b. A person's place of birth
c. The mileage (in miles per gallon) of a car
d. The color of a car

Solution

a. Age is quantitative. It tells how much time has elapsed since the person was born.
b. Place is qualitative. It includes categories such as "New York," "Atlanta," "Denver," and "Los Angeles."
c. Mileage is quantitative. It tells how many miles a car will go on a gallon of gasoline.
d. Color is qualitative. It includes categories such as "blue," "red," and "yellow."

Objective 3 Distinguish between ordinal and nominal variables

Ordinal and Nominal Variables

Qualitative variables come in two types: **ordinal variables** and **nominal variables**. An ordinal variable is one whose categories have a natural ordering. The letter grade received in a class, such as A, B, C, D, or F, is an ordinal variable. A nominal variable is one whose

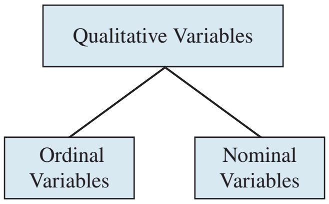


Figure 1.5 Qualitative variables come in two types: ordinal variables and nominal variables.

categories have no natural ordering. Gender is an example of a nominal variable. Figure 1.5 illustrates how qualitative variables are divided into nominal and ordinal variables.

SUMMARY

- Ordinal variables are qualitative variables whose categories have a natural ordering.
- Nominal variables are qualitative variables whose categories have no natural ordering.

EXAMPLE 1.12

Distinguishing between ordinal and nominal variables

Which of the following variables are ordinal and which are nominal?

a. State of residence
b. Gender
c. Letter grade in a statistics class (A, B, C, D, or F)
d. Size of soft drink ordered at a fast-food restaurant (small, medium, or large)

Solution

a. State of residence is nominal. There is no natural ordering to the states.
b. Gender is nominal.
c. Letter grade in a statistics class is ordinal. The order, from high to low, is A, B, C, D, F.
d. Size of soft drink is ordinal.

Objective 4 Distinguish between discrete and continuous variables

Discrete and Continuous Variables

Quantitative variables can be either *discrete* or *continuous*. **Discrete variables** are those whose possible values can be listed. Often, discrete variables result from counting something, so the possible values of the variable are 0, 1, 2, and so forth. **Continuous variables** can, in principle, take on any value within some interval. For example, height is a continuous variable because someone's height can be 68, or 68.1, or 68.1452389 inches. The possible values for height are not restricted to a list. Figure 1.6 illustrates how quantitative variables are divided into discrete and continuous variables.

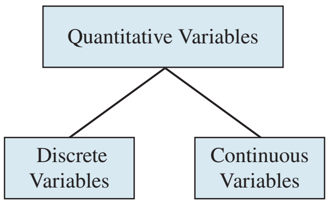


Figure 1.6 Quantitative variables come in two types: discrete variables and continuous variables.

SUMMARY

- Discrete variables are quantitative variables whose possible values can be listed. The list may be infinite—for example, the list of all whole numbers.
- Continuous variables are quantitative variables that can take on any value in some interval. The possible values of a continuous variable are not restricted to any list.

EXAMPLE 1.13

Distinguishing between discrete and continuous variables

Which of the following variables are discrete, and which are continuous?

a. The age of a person at his or her last birthday
b. The height of a person
c. The number of siblings a person has
d. The distance a person commutes to work

Solution

a. Age at a person's last birthday is discrete. The possible values are 0, 1, 2, and so forth.
b. Height is continuous. A person's height is not restricted to any list of values.
c. Number of siblings is discrete. The possible values are 0, 1, 2, and so forth.
d. Distance commuted to work is continuous. It is not restricted to any list of values.

Objective 5 Distinguish between ratio and interval levels of measurement

Ratio and Interval Levels of Measurement

Quantitative variables can be categorized as having a ratio or an interval level of measurement. A variable has a **ratio level of measurement** if a value of zero indicates that none of the quantity is present, and if ratios of values of the variable are meaningful. For example, money is measured at the ratio level. An amount of $0 represents no money. Ratios are meaningful as well: for example, $20 is twice as much as $10.

A variable has the **interval level of measurement** if a value of zero does not indicate that none of the quantity is present. Two commonly encountered interval variables are dates and temperature. There is no value of 0 for dates; we do not denote the beginning of time as the year 0. Similarly, when measured in °F or °C, a temperature of 0° does not indicate an absence of heat. Ratios are not meaningful for interval variables; for example, a temperature of 100° is not twice as hot as a temperature of 50°. Differences are meaningful, however; for example, the year 2030 is 10 years later than the year 2020.

SUMMARY

- A variable has the ratio level of measurement if zero represents the absence of the quantity, and ratios are meaningful.
- A variable has the interval level of measurement if zero does not represent the absence of the quantity, and ratios are not meaningful. Differences are meaningful, however.

EXAMPLE 1.14

Distinguishing between interval and ratio levels of measurement

Which of the following variables are at the interval and which are at the ratio level of measurement?

1. The number of siblings you have
2. The outdoor temperature in °C
3. The year of the next presidential election
4. The price of a pair of shoes

Solution

a. Ratio. Zero siblings means the absence of siblings, and 4 siblings is twice as many as 2.
b. Interval. Temperature in °C has the interval level of measurement, because 0° does not represent the absence of heat.
c. Interval. Dates have the interval level of measurement.
d. Ratio. A price of $0 indicates that the shoes cost no money, and a $100 pair of shoes is twice as expensive as a $50 pair of shoes.

Check Your Understanding

2. Which are qualitative and which are quantitative?
   a. The number of patients admitted to a hospital on a given day
   b. The model of car last sold by a particular car dealer

c. The name of your favorite song
d. The seating capacity of an auditorium

3. Which are nominal and which are ordinal?
a. The names of the streets in a town
b. The movie ratings G, PG, PG-13, R, and NC-17
c. The winners of the gold, silver, and bronze medals in an Olympic swimming competition

4. Which are discrete and which are continuous?
a. The number of female members of the U.S. House of Representatives
b. The amount of water used by a household during a given month
c. The number of stories in an apartment building
d. A person's body temperature

5. Which are at the interval and which are at the ratio level of measurement?
a. The year you started school
b. Your age in years
c. The time that your first class starts
d. The price of a loaf of bread

Answers are on page 19.

SECTION 1.2 Exercises

Exercises 1–5 are the Check Your Understanding exercises located within the section.

Understanding the Concepts

In Exercises 6–12, fill in each blank with the appropriate word or phrase.

6. The characteristics of individuals about which we collect information are called ________.

7. The values of variables are called ________ .

8. Variables that classify individuals into categories are called ________.

9. ________ variables are always numerical.

10. Qualitative variables can be divided into two types: ________ and ________.

11. A ________ variable is a quantitative variable whose possible values can be listed.

12. ________ variables can take on any value in some interval.

In Exercises 13–16, determine whether the statement is true or false. If the statement is false, rewrite it as a true statement.

13. Qualitative variables describe how much or how many of something there is.

14. A nominal variable is a qualitative variable with no natural ordering.

15. A discrete variable is one whose possible values can be listed.

16. A person's height is an example of a continuous variable.

Practicing the Skills

In Exercises 17–26, determine whether the data described are qualitative or quantitative.

17. Your best friend's name

18. Your best friend's age

19. The number of touchdowns in a football game

20. The title of your statistics book

21. The number of files on a computer

22. The waist size of a pair of jeans

23. The ingredients in a recipe

24. Your school colors

25. The makes of cars sold by a particular car dealer

26. The number of cars sold by a car dealer last month

In Exercises 27–34, determine whether the data described are nominal or ordinal.

27. The categories Strongly disagree, Disagree, Neutral, Agree, and Strongly agree on a survey

28. The names of the counties in a state

29. The shirt sizes of Small, Medium, Large, and X-Large

30. I got an A in statistics, a B in biology, and C's in history and English.

31. This semester, I am taking statistics, biology, history, and English.

32. I ordered a pizza with pepperoni, mushrooms, olives, and onions.

33. In the track meet, I competed in the high jump and the pole vault.

34. I finished first in the high jump and third in the pole vault.

In Exercises 35–42, determine whether the data described are discrete or continuous.

35. The amount of caffeine in a cup of Starbucks coffee

36. The distance from a student's home to his school

37. The number of steps in a stairway

38. The number of students enrolled at a college

39. The amount of charge left in a phone battery

40. The number of patients who reported that a new drug had relieved their pain

41. The number of electrical outlets in a coffee shop

42. The time it takes for a text message to be delivered

In Exercises 43–50, determine whether the data described are at the interval level or the ratio level of measurement.

43. The year of your birth

44. The sales price of a car

45. The weight in pounds of a sack of potatoes

46. The score on an SAT exam (range is 200 to 800 points)

47. The water temperature, in °C, in a swimming pool

48. The number of Snapchat friends someone has

49. A student's GPA

50. The amount of data, in gigabytes, remaining on your phone

Working with the Concepts

51. Popular Videos: Following are the ten most-viewed YouTube videos, as of January 2020:

1. Despacito—Luis Fonsi featuring Daddy Yankee
2. Shape of You—Ed Sheeran
3. Baby Shark Dance—Pinkfong Kids' Songs & Stories
4. See You Again—Wiz Khalifa featuring Charlie Puth
5. Masha and the Bear: "Recipe for Disaster"—Get Movies
6. Uptown Funk—Mark Ronson featuring Bruno Mars
7. Gangnam Style—Psy
8. Sorry—Justin Bieber
9. Sugar—Maroon 5
10. Roar—Katy Perry

Source: Wikipedia

a. Are these data nominal or ordinal?
b. Are these data qualitative or quantitative?

52. More Videos: The following table presents the number of views (in millions) for the videos in Exercise 51.

Video	Number of views (millions)
1. Despacito—Luis Fonsi featuring Daddy Yankee	6610
2. Shape of You—Ed Sheeran	4600
3. Baby Shark Dance—Pinkfong Kids' Songs & Stories	4470
4. See You Again—Wiz Khalifa featuring Charlie Puth	4390
5. Masha and the Bear: "Recipe for Disaster"—Get Movies	4220
6. Uptown Funk—Mark Ronson featuring Bruno Mars	3770
7. Gangnam Style—Psy	3500
8. Sorry—Justin Bieber	3240
9. Sugar—Maroon 5	3120
10. Roar—Katy Perry	2990

Source: Wikipedia

a. Are these data discrete or continuous?
b. Are these data qualitative or quantitative?

53. How's the economy? A poll conducted by the American Research Group asked individuals their views on how the economy will be a year from now. Respondents were given four choices: Better than today, Same as today, Worse than today, and Undecided. Are these choices nominal or ordinal?

54. Global warming: A recent Pew poll asked people between the ages of 18 and 29 how serious a problem global warming is. Of those who responded, 43% thought it was very serious, 24% thought it was somewhat serious, 15% thought it was not too serious, and 17% thought it was not a problem. Are these percentages qualitative or quantitative?

55. Graphic Novels: According to *Time* magazine, some of the best graphic novels of all time are:

Watchmen by Alan Moore and Dave Gibbons
Sandman by Neil Gaiman
Jimmy Corrigan, the Smartest Kid on Earth by Chris Ware
Maus by Art Spiegelman
The Adventures of Tintin: The Black Island by Hergé
Miracleman: The Golden Age by Neil Gaiman and Mark Buckingham
Fun Home: A Family Tragicomic by Allison Bechdel
Tha Greatest of Marlys by Lynda Berry

Are these data nominal or ordinal?

56. Watch your language: According to MerriamWebster Online, the top ten Funny Sounding and Interesting words are:

1. Bumfuzzle	6. Snickersnee
2. Cattywampus	7. Widdershins
3. Gardyloo	8. Collywobbles
4. Taradiddle	9. Gubbins
5. Billingsgate	10. Diphthong

Are these data nominal or ordinal?

57. Top ten video games: According to *Wikipedia*, the following are the top ten selling video games of all time.

Game Title	Initial Release Year	Developer	Copies Sold (millions)
1. Tetris	1984	Elektronorgtechnica	170.0
2. Minecraft	2011	Mojang	154.0
3. Grand Theft Auto V	2013	Rockstar North	100.0
4. Wii Sports	2008	Nintendo	82.9
5. PlayerUnknown's Battlegrounds	2017	PUBG Corporation	50.0
6. Pokemon Red/Green/Blue/Yellow	1996	Game Freak	47.5
7. Wii Fit and Wii Fit Plus	2007	Nintendo	43.8
8. Super Mario Bros.	1985	Nintendo	43.2
9. Mario Kart Wii	2008	Wii	37.1
10. Wii Sports Resort	2009	Nintendo	33.1

a. Which of the columns represent qualitative variables?
b. Which of the columns represent quantitative variables?
c. Which of the columns represent nominal variables?
d. Which of the columns represent ordinal variables?

58. At the movies: The following table provides information about the top ten grossing movies of all time.

Movie Title	Release Year	Studio	Ticket Sales (millions of $)	Running Time (minutes)
1. Avengers: Endgame	2019	Disney	2797.8	182
2. Avatar	2009	Fox	2790.4	162
3. Titanic	1997	Paramount	2194.4	194
4. Star Wars: The Force Awakens	2015	Disney	2068.2	136
5. Avengers: Infinity War	2018	Disney	2048.4	160
6. Jurassic World	2015	Univision	1670.4	124
7. The Lion King	2019	Disney	1656.9	118
8. Marvel's The Avengers	2012	Disney	1518.8	143
9. Furious 7	2015	Univision	1515.0	140
10. Frozen II	2019	Disney	1420.8	103

Source: Box Office Mojo

a. Which of the columns represent qualitative variables?
b. Which of the columns represent quantitative variables?
c. Which of the columns represent nominal variables?
d. Which of the columns represent ordinal variables?

Extending the Concepts

59. What do the numbers mean? A survey is administered by a marketing firm. Two of the people surveyed are Brenda and Jason. Three of the questions are as follows:

i. Do you favor the construction of a new shopping mall?
(1) Strongly oppose (2) Somewhat oppose (3) Neutral (4) Somewhat favor (5) Strongly favor

ii. How many cars do you own?

iii. What is your marital status?
(1) Married (2) Single (3) Divorced (4) Domestically partnered (5) Other

a. Are the responses for question (i) nominal or ordinal?
b. On question (i), Brenda answers (2) and Jason answers (4). Jason's answer (4) is greater than Brenda's answer (2). Does Jason's answer reflect more of something?
c. Jason's answer to question (i) is twice as large as Brenda's answer. Does Jason's answer reflect twice as much of something? Explain.
d. Are the responses for question (ii) qualitative or quantitative?

e. On question (ii), Brenda answers 2 and Jason answers 1. Does Brenda's answer reflect more of something? Does Brenda's answer reflect twice as much of something? Explain.
f. Are the responses for question (iii) nominal or ordinal?
g. On question (iii), Brenda answers (4) and Jason answers (2). Does Brenda's answer reflect more of something? Does Brenda's answer reflect twice as much of something? Explain.

Answers to Check Your Understanding Exercises for Section 1.2

1. a. 6 **b.** Political affiliation, Age, and Voted in last election
c. Political affiliation = Democrat, Age = 21, Voted in last election = no

2. a. Quantitative **b.** Qualitative **c.** Qualitative
d. Quantitative

3. a. Nominal **b.** Ordinal **c.** Ordinal

4. a. Discrete **b.** Continuous **c.** Discrete
d. Continuous

5. a. Interval **b.** Ratio **c.** Interval
d. Ratio

SECTION 1.3 Design of Experiments

Objectives

1. Distinguish between a randomized experiment and an observational study
2. Understand the advantages of randomized experiments
3. Understand how confounding can affect the results of an observational study
4. Describe various types of observational studies

Objective 1 Distinguish between a randomized experiment and an observational study

Experiments and Observational Studies

Will a new drug help prevent heart attacks? Does one type of seed produce a larger wheat crop than another? Does exercise lower blood pressure? To illustrate how scientists address questions like these, we describe how a study might be conducted to determine which of three types of seed will result in the largest wheat yield.

- Prepare three identically sized plots of land, with similar soil types.
- Plant each type of seed on a different plot, choosing the plots at random.
- Water and fertilize the plots in the same way.
- Harvest the wheat, and measure the amount grown on each plot.
- If one type of seed produces substantially more (or less) wheat than the others, then scientists will conclude that it is better (or worse) than the others.

The following terminology is used for studies like this.

DEFINITION

The **experimental units** are the individuals that are studied. These can be people, animals, plants, or things. When the experimental units are people, they are sometimes called **subjects**.

In the wheat study just described, the experimental units are the three plots of land.

DEFINITION

The **outcome**, or **response**, is what is measured on each experimental unit.

In the wheat study, the outcome is the amount of wheat produced.

DEFINITION

The **treatments** are the procedures applied to each experimental unit. There are always two or more treatments. The purpose is to determine whether the choice of treatment affects the outcome.

In the wheat study, the treatments are the three types of seed.

In general, studies fall into two categories: *randomized experiments* and *observational studies*.

DEFINITION

A **randomized experiment** is a study in which the investigator assigns the treatments to the experimental units at random.

The wheat study described above is a randomized experiment. In some situations, randomized experiments cannot be performed, because it isn't possible to randomly assign the treatments. For example, in studies to determine how smoking affects health, people cannot be assigned to smoke. Instead, people choose for themselves whether to smoke, and scientists observe differences in health outcomes between groups of smokers and nonsmokers. Studies like this are called *observational studies*.

DEFINITION

An **observational study** is one in which the assignment to treatment groups is not made by the investigator.

When possible, it is better to assign treatments at random and perform a randomized experiment. As we will see, the results of randomized experiments are generally easier to interpret than the results of observational studies.

Objective 2 Understand the advantages of randomized experiments

Randomized Experiments

Ingram Publishing/SuperStock

An article in *The New England Journal of Medicine* (359:339–354) reported the results of a study to determine whether a drug called raltegravir is effective in reducing levels of virus in patients with human immunodeficiency virus (HIV). A total of 699 patients participated in the experiment. These patients were divided into two groups. One group was given raltegravir. The other group was given a placebo. (A placebo is a harmless tablet, such as a sugar tablet, that looks like the drug but has no medical effect.) Thus there were two treatments in this experiment, raltegravir and placebo.

The experimenters had decided to give raltegravir to about two-thirds of the subjects and the placebo to the others. To determine which patients would be assigned to which group, a simple random sample consisting of 462 of the 699 patients was drawn; this sample constituted the raltegravir group. The remaining 237 patients were assigned to the placebo group.

It was decided to examine subjects after 16 weeks and measure the levels of virus in their blood. Thus the outcome for this experiment was the number of copies of virus per milliliter of blood. Patients were considered to have a successful outcome if they had fewer than 50 copies of the virus per milliliter of blood. In the raltegravir group, 62% of the subjects had a successful outcome, but only 35% of the placebo group did. The conclusion was that raltegravir was effective in lowering the concentration of virus in HIV patients. We will examine this study and determine why it was reasonable to reach this conclusion.

The raltegravir study was a randomized experiment, because the treatments were assigned to the patients at random. What are the advantages of randomized experiments? In a perfect study, the treatment groups would not differ from each other in any important way except that they receive different treatments. Then, if the outcomes differ among the groups, we may be confident that the differences in outcome must have been caused by differences in treatment. In practice, it is impossible to construct treatment groups that are exactly alike. But randomization does the next best thing. In a randomized experiment, any differences between the groups are likely to be small. In addition, the differences are due only to chance.

Because the raltegravir study was a randomized experiment, it is reasonable to conclude that the higher success rate in the raltegravir group was actually due to raltegravir.

SUMMARY

In a randomized experiment, if there are large differences in outcomes among the treatment groups, we can conclude that the differences are due to the treatments.

EXAMPLE 1.15

Identifying a randomized experiment

To assess the effectiveness of a new method for teaching arithmetic to elementary school children, a simple random sample of 30 first graders was taught with the new method, and another simple random sample of 30 first graders was taught with the currently used method. At the end of eight weeks, the children were given a test to assess their knowledge. What are the treatments in this study? Explain why this is a randomized experiment.

Solution

The treatments are the two methods of teaching. This is a randomized experiment because children were assigned to the treatment groups at random.

Double-blind experiments

We have described the advantages of assigning treatments at random. It is a further advantage if the assignment can be done in such a way that neither the experimenters nor the subjects know which treatment has been assigned to which subject. Experiments like this are called *double-blind* experiments. The raltegravir experiment was a double-blind experiment, because neither the patients nor the doctors treating them knew which patients were receiving the drug and which were receiving the placebo.

DEFINITION

An experiment is **double-blind** if neither the investigators nor the subjects know who has been assigned to which treatment.

Experiments should be run double-blind whenever possible, because when investigators or subjects know which treatment is being given, they may tend to report the results differently. For example, in an experiment to test the effectiveness of a new pain reliever, patients who know they are getting the drug may report their pain levels differently than those who know they are taking a placebo. Doctors can be affected as well; a doctor's diagnosis may be influenced by a knowledge of which treatment a patient received.

In some situations, it is not possible to run a double-blind experiment. For example, in an experiment that compares a treatment that involves taking medication to a treatment that involves surgery, both patients and doctors will know who got which treatment.

EXAMPLE 1.16

Determining whether an experiment is double-blind

Is the experiment described in Example 1.15 a double-blind experiment? Explain.

Solution

This experiment is not double-blind, because the teachers know whether they are using the new method or the old method.

Randomized block experiments

The type of randomized experiment we have discussed is sometimes called a **completely randomized experiment**, because there is no restriction on which subjects may be assigned which treatment. In some situations, it is desirable to restrict the randomization a bit. For example, imagine that two reading programs are to be tested in an elementary school that has children in grades 1 through 4. If children are assigned at random to the programs, it is possible that one of the programs will end up with more fourth graders while the other one will end up with more first graders. Since fourth graders tend to be better readers, this will give an advantage to the program that happens to end up with more of them. This possibility can be avoided by randomizing the students within each grade separately. In other words, we randomly assign exactly half of the students within each grade to each reading program.

This type of experiment is called a **randomized block experiment**. In the example just discussed, each grade constitutes a block. In a randomized block experiment, the subjects are divided into blocks in such a way that the subjects in each block are the same or similar with regard to a variable that is related to the outcome. Age and gender are commonly used blocking variables. Then the subjects within each block are randomly assigned a treatment.

Observational Studies

Recall that an observational study is one in which the investigators do not assign the treatments. In most observational studies, the subjects choose their own treatments. Observational studies are less reliable than randomized experiments. To see why, imagine a study that is intended to determine whether smoking increases the risk of heart attack. Imagine that a group of smokers and a group of nonsmokers are observed for several years, and during that time a higher percentage of the smoking group experiences a heart attack. Does this prove that smoking increases the risk of heart attack? No. The problem is that the smoking group will differ from the nonsmoking group in many ways other than smoking, and these other differences may be responsible for differences in the rate of heart attacks. For example, smoking is more prevalent among men than among women. Therefore, the smoking group will contain a higher percentage of men than the nonsmoking group. It is known that men have a higher risk of heart attack than women. So the higher rate of heart attacks in the smoking group could be due to the fact that there are more men in the smoking group, and not to the smoking itself.

Objective 3 Understand how confounding can affect the results of an observational study

Confounding

The preceding example illustrates the major problem with observational studies. It is difficult to tell whether a difference in the outcome is due to the treatment or to some other difference between the treatment and control groups. This is known as **confounding**. In the preceding example, gender was a *confounder*. Gender is related to smoking (men are more likely to smoke) and to heart attacks (men are more likely to have heart attacks). For this reason, it is difficult to determine whether the difference in heart attack rates is due to differences in smoking (the treatment) or differences in gender (the confounder).

Explain It Again

Another way to describe a confounder: A confounder is something other than the treatment that can cause the treatment groups to have different outcomes.

SUMMARY

A **confounder** is a variable that is related to both the treatment and the outcome. When a confounder is present, it is difficult to determine whether differences in the outcome are due to the treatment or to the confounder.

How can we prevent confounding? One way is to design a study so that the confounder isn't a factor. For example, to determine whether smoking increases the risk of heart attack, we could compare a group of male smokers to a group of male nonsmokers, and a group of female smokers to a group of female nonsmokers. Gender wouldn't be a confounder here, because there would be no differences in gender between the smoking and nonsmoking groups. Of course, there are other possible confounders. Smoking rates vary among ethnic groups, and rates of heart attacks do, too. If people in ethnic groups that are more susceptible to heart attacks are also more likely to smoke, then ethnicity becomes a confounder. This can be dealt with by comparing smokers of the same gender and ethnic group to nonsmokers of that gender and ethnic group.

Designing observational studies that are relatively free of confounding is difficult. In practice, many studies must be conducted over a long period of time. In the case of smoking, this has been done, and we can be confident that smoking does indeed increase the risk of heart attack, along with other diseases. If you don't smoke, you have a much better chance to live a long and healthy life.

SUMMARY

In an observational study, when there are differences in the outcomes among the treatment groups, it is often difficult to determine whether the differences are due to the treatments or to confounding.

EXAMPLE 1.17

Determining the effect of confounding

In a study of the effects of blood pressure on health, a large group of people of all ages were given regular blood pressure checkups for a period of one year. It was found that people with high blood pressure were more likely to develop cancer than people with lower blood pressure. Explain how this result might be due to confounding.

Solution

Age is a likely confounder. Older people tend to have higher blood pressure than younger people, and older people are more likely to get cancer than younger people. Therefore people with high blood pressure may have higher cancer rates than younger people, even though high blood pressure does not cause cancer.

Check Your Understanding

1. To study the effect of air pollution on respiratory health, a group of people in a city with high levels of air pollution and another group in a rural area with low levels of pollution are examined to determine their lung capacity. Is this a randomized experiment or an observational study?
2. It is known that drinking alcohol increases the risk of contracting liver cancer. Assume that in an observational study, a group of smokers has a higher rate of liver cancer than a group of nonsmokers. Explain how this result might be due to confounding.

Answers are on page 27.

Objective 4 Describe various types of observational studies

Types of Observational Studies

There are two main types of observational studies: cohort studies and case-control studies. Cohort studies can be further divided into prospective, cross-sectional, and retrospective studies.

Cohort studies

In a **cohort study**, a group of subjects (the cohort) is studied to determine whether various factors of interest are associated with an outcome.

In a **prospective** cohort study, the subjects are followed over time. One of the most famous prospective cohort studies is the Framingham Heart Study. This study began in 1948 with 5209 men and women from the town of Framingham, Massachusetts. Every two years, these subjects are given physical exams and lifestyle interviews, which are studied to discover factors that increase the risk of heart disease. Much of what is known about the effects of diet and exercise on heart disease is based on this study.

Prospective studies are among the best observational studies. Because subjects are repeatedly examined, the quality of the data is often quite good. Information on potential confounders can be collected as well. Results from prospective studies are generally more reliable than those from other observational studies. The disadvantages of prospective studies are that they are expensive to run and that it takes a long time to develop results.

In a **cross-sectional** study, measurements are taken at one point in time. For example, in a study published in the *Journal of the American Medical Association* (300:1303–1310), I. Lang and colleagues studied the health effects of bisphenol A, a chemical found in the linings of food and beverage containers. They measured the levels of bisphenol A in urine samples from 1455 adults. They found that people with higher levels of bisphenol A were more likely to have heart disease and diabetes.

Cross-sectional studies are relatively inexpensive, and results can be obtained quickly. The main disadvantage is that the exposure is measured at only one point in time, so there is little information about how past exposures may have contributed to the outcome. Another disadvantage is that because measurements are made at only one time, it is impossible to determine a time sequence of events. For example, in the bisphenol A study just described, it is possible that higher levels of bisphenol A cause heart disease and diabetes. But it is also possible that the onset of heart disease or diabetes causes levels of bisphenol A to increase. There is no way to determine which happened first.

In a **retrospective** cohort study, subjects are sampled after the outcome has occurred. The investigators then look back over time to determine whether certain factors are related to the outcome. For example, in a study published in *The New England Journal of Medicine* (357:753–761), T. Adams and colleagues sampled 9949 people who had undergone gastric bypass surgery between 5 and 15 years previously, along with 9668 obese patients who had not had bypass surgery. They looked back in time to see which patients were still alive. They found that the survival rates for the surgery patients were greater than for those who had not undergone surgery.

Retrospective cohort studies are less expensive than prospective cohort studies, and results can be obtained quickly. A disadvantage is that it is often impossible to obtain data on potential confounders.

One serious limitation of all cohort studies is that they cannot be used to study rare diseases. Even in a large cohort, very few people will contract a particular rare disease. To study rare diseases, case-control studies must be used.

Case-control studies

In a **case-control** study, two samples are drawn. One sample consists of people who have the disease of interest (the cases), and the other consists of people who do not have the disease (the controls). The investigators look back in time to determine whether a particular factor of interest differs between the two groups. For example, S. S. Nielsen and colleagues conducted a case-control study to determine whether exposure to pesticides is related to brain cancer in children (*Environmental Health Perspectives*, 118:144–149). They sampled 201 children under the age of 10 who had been diagnosed with brain cancer, and 285 children who did not have brain cancer. They interviewed the parents of the children to estimate the extent to which the children had been exposed to pesticides. They did not find a clear relationship between pesticide exposure and brain cancer. This study could not have been conducted as a cohort study, because even in a large cohort of children, very few will get brain cancer.

Case-control studies are always retrospective, because the outcome (case or control) has occurred before the sampling is done. Case-control studies have the same advantages and disadvantages as retrospective cohort studies. In addition, case-control studies have the advantage that they can be used to study rare diseases.

Check Your Understanding

3. In a study conducted at the University of Southern California, J. Peters and colleagues studied elementary school students in 12 California communities. Each year for 10 years, they measured the respiratory function of the children and the levels of air pollution in the communities.

a. Was this a cohort study or a case-control study?

b. Was the study prospective, cross-sectional, or retrospective?

4. In a study conducted at the University of Colorado, J. Ruttenber and colleagues studied people who had worked at the Rocky Flats nuclear weapons production facility near Denver, Colorado. They studied a group of workers who had contracted lung cancer and another group who had not contracted lung cancer. They looked back at plant records to determine the amount of radiation exposure for each worker. The purpose of the study was to determine whether the people with lung cancer had been exposed to higher levels of radiation than those who had not gotten lung cancer.
a. Was this a cohort study or a case-control study?
b. Was the study prospective, cross-sectional, or retrospective?

Answers are on page 27.

SECTION 1.3 Exercises

Exercises 1–4 are the Check Your Understanding exercises located within the section.

Understanding the Concepts

In Exercises 5–10, fill in each blank with the appropriate word or phrase.

5. In a __________ experiment, subjects do not decide for themselves which treatment they will get.

6. In a __________ study, neither the investigators nor the subjects know who is getting which treatment.

7. A study in which the assignment to treatment groups is not made by the investigator is called __________.

8. A __________ is a variable related to both the treatment and the outcome.

9. In a __________ study, the subjects are followed over time.

10. In a __________ study, a group of subjects is studied to determine whether various factors of interest are associated with an outcome.

In Exercises 11–16, determine whether the statement is true or false. If the statement is false, rewrite it as a true statement.

11. In a randomized experiment, the treatment groups do not differ in any systematic way except that they receive different treatments.

12. A confounder makes it easier to draw conclusions from a study.

13. In an observational study, subjects are assigned to treatment groups at random.

14. Observational studies are generally more reliable than randomized experiments.

15. In a case-control study, the outcome has occurred before the subjects are sampled.

16. In a cross-sectional study, measurements are made at only one point in time.

Practicing the Skills

17. To determine the effectiveness of a new pain reliever, a randomly chosen group of pain sufferers is assigned to take the new drug, and another randomly chosen group is assigned to take a placebo.
a. Is this a randomized experiment or an observational study?
b. The subjects taking the new drug experienced substantially more pain relief than those taking the placebo. The research team concluded that the new drug is effective in relieving pain. Is this conclusion well justified? Explain.

18. A medical researcher wants to determine whether exercising can lower blood pressure. At a health fair, he measures the blood pressure of 100 individuals and interviews them about their exercise habits. He divides the individuals into two categories: those whose typical level of exercise is low, and those whose level of exercise is high.
a. Is this a randomized experiment or an observational study?
b. The subjects in the low-exercise group had considerably higher blood pressure, on the average, than subjects in the high-exercise group. The researcher concluded that exercise decreases blood pressure. Is this conclusion well justified? Explain.

19. A medical researcher wants to determine whether exercising can lower blood pressure. She recruits 100 people with high blood pressure to participate in the study. She assigns a random sample of 50 of them to pursue an exercise program that includes daily swimming and jogging. She assigns the other 50 to refrain from vigorous activity. She measures the blood pressure of each of the 100 individuals both before and after the study.
a. Is this a randomized experiment or an observational study?
b. On the average, the subjects in the exercise group substantially reduced their blood pressure, while the subjects in the no-exercise group did not experience a reduction. The researcher concluded that exercise decreases blood pressure. Is this conclusion well justified? Explain.

20. An agricultural scientist wants to determine the effect of fertilizer type on the yield of tomatoes. There are four types of fertilizer under consideration. She plants tomatoes on four plots of land. Each plot is treated identically except for receiving a different type of fertilizer.
a. What are the treatments?
b. Is this a randomized experiment or an observational study?
c. The yields differ substantially among the four plots. Can you conclude that the differences in yield are due to the differences in fertilizer? Explain.

hüseyin harmandaẑlž/Getty Images

Working with the Concepts

21. Air pollution and colds: A scientist wants to determine whether people who live in places with high levels of air pollution get more colds than people in areas with little air pollution. Do you think it is possible to design a randomized experiment to study this question, or will an observational study be necessary? Explain.

22. Cold medications: A scientist wants to determine whether a new cold medicine relieves symptoms more effectively than a currently used medicine. Do you think it is possible to design a randomized experiment to study this question, or will an observational study be necessary? Explain.

23. Taxicabs and crime: A sociologist discovered that regions that have more taxicabs tend to have higher crime rates. Does increasing the number of taxicabs cause the crime rate to increase, or could the result be due to confounding? Explain.

24. Recovering from heart attacks: In a study of people who had suffered heart attacks, it was found that those who lived in smaller houses were more likely to recover than those who lived in larger houses. Does living in a smaller house increase the likelihood of recovery from a heart attack, or could the result be due to confounding? Explain.

25. Eat your vegetables: In an observational study, people who ate four or more servings of fresh fruits and vegetables each day were less likely to develop colon cancer than people who ate little fruit or vegetables. True or false:

a. The results of the study show that eating more fruits and vegetables reduces your risk of contracting colon cancer.

b. The results of the study may be due to confounding, since the lifestyles of people who eat large amounts of fruits and vegetables may differ in many ways from those of people who do not.

26. Vocabulary and height: A vocabulary test was given to students at an elementary school. The students' ages ranged from 5 to 11 years old. It was found that the students with larger vocabularies tended to be taller than the students with smaller vocabularies. Explain how this result might be due to confounding.

27. Secondhand smoke: A recent study compared the heart rates of 19 infants born to nonsmoking mothers with those of 17 infants born to mothers who smoked an average of 15 cigarettes a day while pregnant and after giving birth. The heart rates of the infants at one year of age were 20% slower on the average for the smoking mothers.

a. What is the outcome variable?

b. What is the treatment variable?

c. Was this a cohort study or a case-control study?

d. Was the study prospective, cross-sectional, or retrospective?

e. Could the results be due to confounding? Explain.

Source: *Environmental Health Perspectives* 118:a158–a159

28. Pollution in China: In a recent study, Z. Zhao and colleagues measured the levels of formaldehyde in the air in 34 classrooms in the schools in the city of Taiyuan, China. On the same day, they gave questionnaires to 1993 students aged 11–15 in those schools, asking them whether they had experienced respiratory problems (such as asthma attacks, wheezing, or shortness of breath). They found that the students in the classrooms with higher levels of formaldehyde reported more respiratory problems.

a. What is the outcome variable?

b. What is the treatment variable?

c. Was this a cohort study or a case-control study?

d. Was the study prospective, cross-sectional, or retrospective?

e. Could the results be due to confounding? Explain.

Source: *Environmental Health Perspectives* 116:90–97

Extending the Concepts

29. The Salk Vaccine Trial: In 1954, the first vaccine against polio, known as the Salk vaccine, was tested in a large randomized double-blind study. Approximately 750,000 children were asked to enroll in the study. Of these, approximately 350,000 did not participate, because their parents refused permission. The children who did participate were randomly divided into two groups of about 200,000 each. One group, the treatment group, got the vaccine, while the other group, the control group, got a placebo. The rate of polio in the treatment group was less than half of that in the control group.

a. Is it reasonable to conclude that the Salk vaccine was effective in reducing the rate of polio?

b. Polio is sometimes difficult to diagnose, as its early symptoms are similar to those of the flu. Explain why it was important for the doctors in the study not to know which children were getting the vaccine.

c. Perhaps surprisingly, polio was more common among upper-income and middle-income children than among lower-income children. The reason is that lower-income children tended to live in less hygienic surroundings. They would contract mild cases of polio in infancy while still protected by their mother's antibodies, and thereby develop a resistance to the disease. The children who did not participate in the study were more likely to come from lower-income families. The rate of polio in this group was substantially lower than the rate in the placebo group. Does this prove that the placebo caused polio, or could this be due to confounding? Explain.

30. Another Salk Vaccine Trial: Another study of the Salk vaccine, conducted at the same time as the trial described in