Cover Design: Wiley
Cover Images: Binary code abstract tunnel © loops7/Getty Images,
Data Vortex Tunnel Photography © MirageC/Getty Images

www.wiley.com

WILEY

Wiley Loose-Leaf Print Edition

THIRD EDITION

STATISTICS

UNLOCKING THE POWER OF DATA

ROBIN H. LOCK • PATTI FRAZER LOCK • KARI LOCK MORGAN • ERIC F. LOCK • DENNIS F. LOCK

ISBN 978-1-119-68216-5
90000

WILEY

Statistics

UNLOCKING THE POWER OF DATA

THIRD EDITION

Statistics

UNLOCKING THE POWER OF DATA

THIRD EDITION

Robin H. Lock
St. Lawrence University

Patti Frazer Lock
St. Lawrence University

Kari Lock Morgan
Pennsylvania State University

Eric F. Lock
University of Minnesota

Dennis F. Lock
Buffalo Bills NFL Franchise

WILEY

EXECUTIVE EDITOR	Joanna Dingle & Andrew Noble
EDITORIAL ASSISTANT	Hannah Larkin
SENIOR COURSE CONTENT DEVELOPER	Andrew Moore
SENIOR MANAGER, COURSE DEVELOPMENT AND PRODUCTION	Svetlana Barskaya
SENIOR COURSE PRODUCTION OPERATIONS SPECIALIST	Laura Abrams
MARKETING MANAGER	Michael MacDougald
DESIGNER	Thomas Nery
PHOTO EDITOR	Deepthi Saseendran
COVER DESIGN	Thomas Nery
COVER PHOTO	© Binary Code Tunnel: Getty Images / loops7 Data Vortex Tunnel: Getty Images / MirageC

This book was typeset in 10/12 TimesTen at Lumina Datamatics.

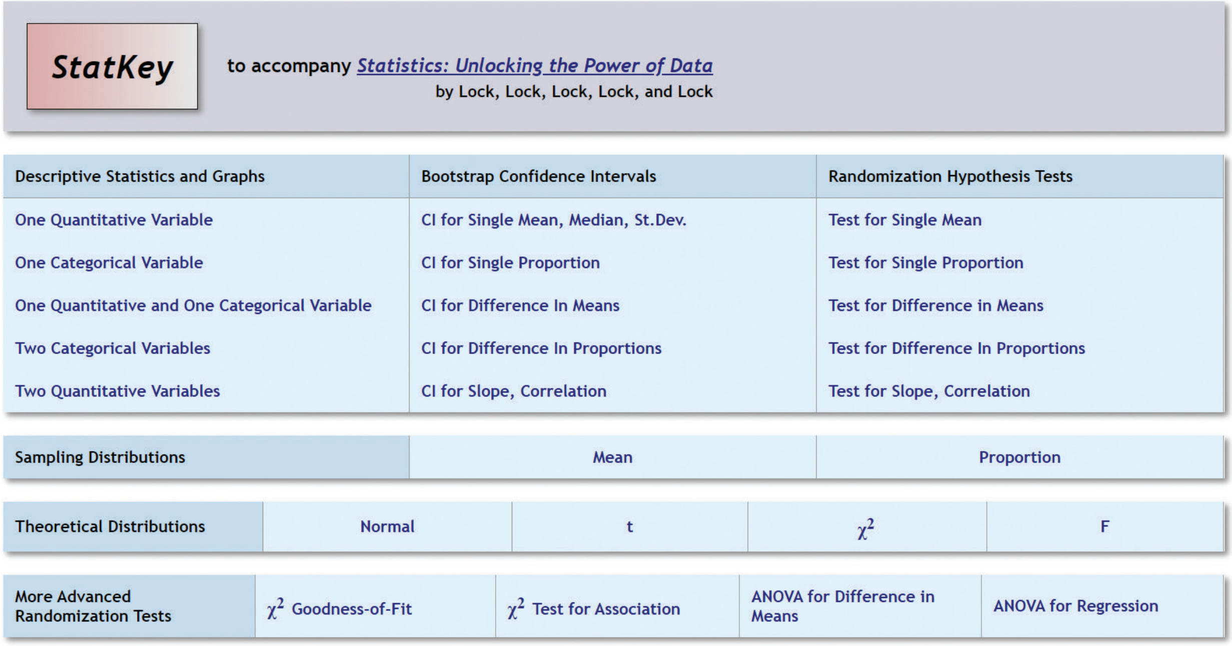
ISBN: 978-1-119-67416-0

Library of Congress Cataloging-in-Publication Data

Names: Lock, Robin H., author. | Lock, Patti Frazer, author. | Morgan,
Kari Lock, author. | Lock, Eric F., author. | Lock, Dennis F., author.
Title: Statistics : Unlocking the Power of Data / Robin H. Lock, St. Lawrence
University, Patti Frazer Lock, St. Lawrence University, Kari Lock Morgan,
Pennsylvania State University, Eric F. Lock, University of Minnesota,
Dennis F. Lock, Buffalo Bills NFL Franchise.
Description: Third edition. | Hoboken : Wiley, 2020. | Originally published:
Statistics / Robin H. Lock ... [et al.].
Identifiers: LCCN 2020036199 (print) | LCCN 2020036200 (ebook) | ISBN 9781119682165
(paperback) | ISBN 9781119682219 (adobe pdf) | ISBN 9781119674160 (epub)
Subjects: LCSH: Statistics–Textbooks.
Classification: LCC QA276.12 .S738 2021 (print) | LCC QA276.12 (ebook) | DDC 519.5–dc23
LC record available at https://lccn.loc.gov/2020036199
LC ebook record available at https://lccn.loc.gov/2020036200

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1



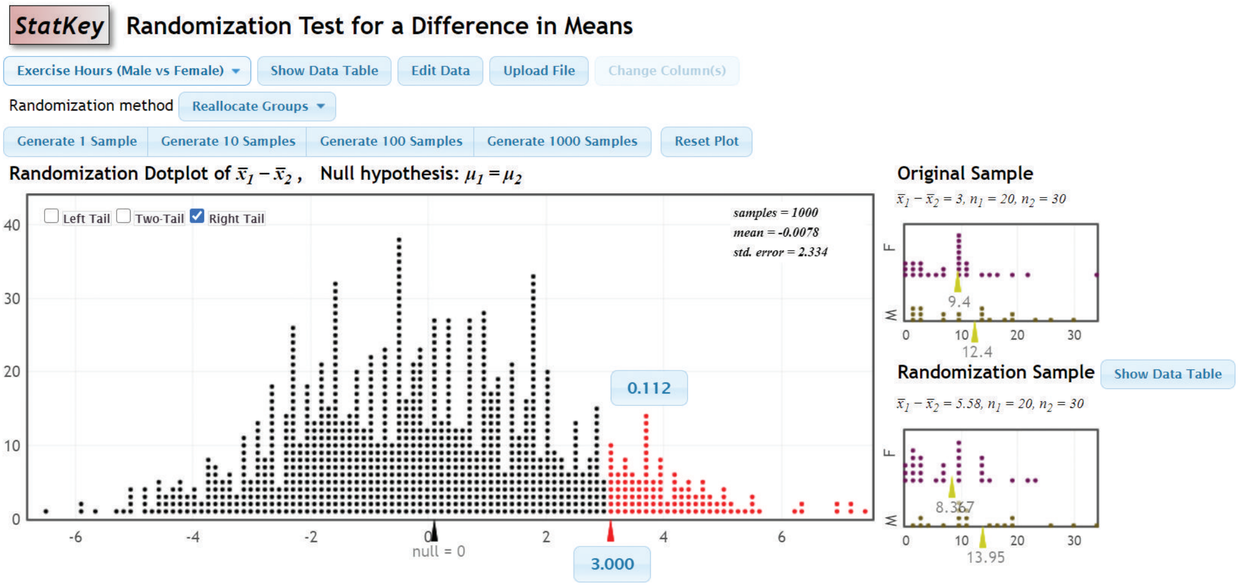
StatKey to accompany *Statistics: Unlocking the Power of Data*
by Lock, Lock, Lock, Lock, and Lock

Descriptive Statistics and Graphs	Bootstrap Confidence Intervals	Randomization Hypothesis Tests
One Quantitative Variable	CI for Single Mean, Median, St.Dev.	Test for Single Mean
One Categorical Variable	CI for Single Proportion	Test for Single Proportion
One Quantitative and One Categorical Variable	CI for Difference In Means	Test for Difference in Means
Two Categorical Variables	CI for Difference In Proportions	Test for Difference In Proportions
Two Quantitative Variables	CI for Slope, Correlation	Test for Slope, Correlation

Sampling Distributions	Mean	Proportion

Theoretical Distributions	Normal	t	χ^2	F

More Advanced Randomization Tests	χ^2 Goodness-of-Fit	χ^2 Test for Association	ANOVA for Difference in Means	ANOVA for Regression



StatKey Randomization Test for a Difference in Means

Exercise Hours (Male vs Female)

Randomization method Reallocate Groups

Randomization Dotplot of $\bar{x}_1 - \bar{x}_2$, Null hypothesis: $\mu_1 = \mu_2$

□ Left Tail □ Two-Tail ☑ Right Tail

samples = 1000
mean = -0.0078
std. error = 2.334

Original Sample

$\bar{x}_1 - \bar{x}_2 = 3, n_1 = 20, n_2 = 30$

Randomization Sample

$\bar{x}_1 - \bar{x}_2 = 5.58, n_1 = 20, n_2 = 30$

You will find *StatKey* and many additional resources
(including short, helpful videos for all examples and all
learning goals; electronic copies of all datasets; and
technology help for a wide variety of platforms)
as part of the online companion to this textbook.

You can also find *StatKey* at

www.lock5stat.com/statkey

A message from the Locks

Data are everywhere—in vast quantities, spanning almost every topic. Being able to make sense of all this information is becoming both a coveted and necessary skill. This book will help you learn how to effectively collect and analyze data, enabling you to investigate any questions you wish to ask. The goal of this book is to help you unlock the power of data!

An essential component of statistics is ***randomness***. Rather than viewing randomness as a confused jumble of numbers (as the random number table on the front cover might appear), you will learn how to use randomness to your advantage, and come to view it as one of the most powerful tools available for making new discoveries and bringing clarity to the world.

PREFACE

"Statistical thinking will one day be as necessary a qualification for efficient citizenship as the ability to read and write."

H.G. Wells

Why We Wrote this Book

Helping students make sense of data will serve them well in life and in any field they might choose. Our goal in writing this book is to help students understand, appreciate, and use the power of statistics and to help instructors teach an outstanding course in statistics.

The text is designed for use in an introductory statistics course. The focus throughout is on data analysis and the primary goal is to enable students to effectively collect data, analyze data, and interpret conclusions drawn from data. The text is driven by real data and real applications. Although the only prerequisite is a minimal working knowledge of algebra, students completing the course should be able to accurately interpret statistical results and to analyze straightforward datasets. The text is designed to give students a sense of the power of data analysis; our hope is that many students learning from this book will want to continue developing their statistical knowledge.

Students who learn from this text should finish with

- A solid conceptual understanding of the key concepts of statistical inference: estimation with intervals and testing for significance.
- The ability to do straightforward data analysis, including summarizing data, visualizing data, and inference using either traditional methods or modern resampling methods.
- Experience using technology to perform a variety of different statistical procedures.
- An understanding of the importance of data collection, the ability to recognize limitations in data collection methods, and an awareness of the role that data collection plays in determining the scope of inference.
- The knowledge of which statistical methods to use in which situations and the ability to interpret the results effectively and in context.
- An awareness of the power of data.

Building Conceptual Understanding with Simulation Methods

This book uses computer simulation methods to introduce students to the key ideas of statistical inference. Methods such as bootstrap intervals and randomization tests are very intuitive to novice students and capitalize on visual learning skills students bring to the classroom. With proper use of computer support, they are accessible at very early stages of a course with little formal background. Our text introduces statistical inference through these resampling and randomization methods, not only because these methods are becoming increasingly important for statisticians in their own right but also because they are outstanding in building students' conceptual understanding of the key ideas.

Our text includes the more traditional methods such as t-tests, chi-square tests, etc., but only after students have developed a strong intuitive understanding of inference through randomization methods. At this point students have a conceptual understanding and appreciation for the results they can then compute using the more traditional methods. We believe that this approach helps students realize that although the formulas may take different forms for different types of data, the conceptual framework underlying most statistical methods remains the same. Our experience has been that after using the intuitive simulation-based methods to introduce the core ideas, students understand and can move quickly through most of the traditional techniques.

Sir R.A. Fisher, widely considered the father of modern statistics, said of simulation and permutation methods in 1936:

"Actually, the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method."

Modern technology has made these methods, too 'tedious' to apply in 1936, now readily accessible. As George Cobb wrote in 2007:

"... despite broad acceptance and rapid growth in enrollments, the consensus curriculum is still an unwitting prisoner of history. What we teach is largely the technical machinery of numerical approximations based on the normal distribution and its many subsidiary cogs. This machinery was once necessary, because the conceptually simpler alternative based on permutations was computationally beyond our reach. Before computers statisticians had no choice. These days we have no excuse. Randomization-based inference makes a direct connection between data production and the logic of inference that deserves to be at the core of every introductory course."

Building Understanding and Proficiency with Technology

Technology is an integral part of modern statistics, but this text does not require any specific software. We have developed a user-friendly set of online interactive dynamic tools, ***StatKey***, to illustrate key ideas and analyze data with modern simulation-based methods. *StatKey* is freely available with data from the text integrated. We also provide **Companion Manuals**, tied directly to the text, for other popular technology options. The text uses many real datasets which are electronically available in multiple formats.

Building a Framework for the Big Picture: Essential Synthesis

One of the drawbacks of many current texts is the fragmentation of ideas into disjoint pieces. While the segmentation helps students understand the individual pieces, we believe that integration of the parts into a coherent whole is also essential. To address this we have sections called Essential Synthesis at the end of each unit, in which students are asked to take a step back and look at the big picture. We hope that these sections, which include case studies, will help to prepare students for the kind of statistical thinking they will encounter after finishing the course.

Building Student Interest with Engaging Examples and Exercises

This text contains over 300 fully worked-out examples and over 2000 exercises, which are the heart of this text and the key to learning statistics. One of the great things about statistics is that it is relevant in so many fields. We have tried to find studies and datasets that will capture the interest of students—and instructors! We hope all users of this text find many fun and useful tidbits of information from the datasets, examples, and exercises, above and beyond the statistical knowledge gained.

The exercise sets at the end of every section assess computation, interpretation, and understanding using a variety of problem types. Some features of the exercise sets include:

- *Skill Builders.* After every section, the exercise set starts with skill-building exercises, designed to be straightforward and to ensure that students have the basic skills and confidence to tackle the more involved problems with real data.
- *Lots of real data.* After the opening skill builders, the vast majority of the exercises in a section involve real data from a wide variety of disciplines. These allow students to practice the ideas of the section and to see how statistics is used in actual practice in addition to illustrating the power and wide applicability of statistics. Most of these exercises call for interpretations of the statistical findings in the context of a real situation.
- *Exercises using technology.* While many exercises provide summary statistics, some problems in each exercise set invite students to use technology to analyze raw data. All datasets, and software-specific companion manuals, are available electronically.
- *Essential synthesis and review.* Exercises at the end of each unit let students choose from among a wider assortment of possible approaches, without the guiding cues associated with section-specific exercise sets. These exercises help students see the big picture and prepare them for determining appropriate analysis methods.

Building Confidence with Robust Student and Instructor Resources

This text has many additional resources designed to facilitate and enhance its use in teaching and learning statistics. The following are all readily accessible and organized to make them easy to find and easy to use. Almost all were written exclusively by the authors.

Resources for students and instructors:

- *StatKey*; online interactive dynamic tools (*www.lock5stat.com/statkey*)
- Software-specific companion manuals (*www.wiley.com/college/lock*)
- All datasets in multiple formats (*www.wiley.com/college/lock*)
- Short video solutions for all examples and video tutorials for all learning goals
- WileyPLUS—an innovative, research-based online environment for effective teaching and learning
- Student solution manual with fully worked solutions to odd-numbered exercises
- Interactive video lectures for every section

Resources for instructors

- Complete instructors manual with sample syllabi, teaching tips and recommended class examples, class activities, homework assignments, and student project assignments
- Short videos with teaching tips for instructors, for every section

- Detailed interactive class activities with handouts, for every section
- PowerPoint slides, for every section, with or without integrated clicker questions
- In-class worksheets ready to go, for every section
- Clicker questions, for every section
- A variety of different types of student projects, for every unit
- Fully worked out solutions to all exercises
- TestGen computerized test bank with a wide variety of provided questions as well as the ability to write custom questions
- The full WileyPLUS learning management system at your disposal

Content and Organization

UNIT A: Data

The first unit deals with data—how to obtain data (Chapter 1) and how to summarize and visualize the information in data (Chapter 2). We explore how the method of data collection influences the types of conclusions that can be drawn and how the type of data (categorical or quantitative) helps determine the most appropriate numerical and/or graphical technique for describing a single variable or investigating a relationship between two variables. We end the unit discussing multiple variables and exploring a variety of additional ways to display data.

UNIT B: Understanding Inference

In Unit B we develop the key ideas of statistical inference—estimation and testing—using simulation methods to build understanding and to carry out the analysis. Chapter 3 introduces the idea of using information from a sample to provide an estimate for a population, and uses a bootstrap distribution to determine the uncertainty in the estimate. In Chapter 4 we illustrate the important ideas for testing statistical hypotheses, again using simple simulations that mimic the random processes of data production. A key feature of these simulation methods is that the general ideas can be applied to a wide variety of parameters and situations.

UNIT C: Inference with Normal and t-Distributions

In Unit C we see how theoretical distributions, such as the classic, bell-shaped normal curve, can be used to approximate the distributions of sample statistics that we encounter in Unit B. Chapter 5 shows, in general, how the normal curve can be used to facilitate constructing confidence intervals and conducting hypothesis tests. In Chapter 6 we see how to estimate standard errors with formulas and use the normal or t-distributions in situations involving means, proportions, differences in means, and differences in proportions. Since the main ideas of inference have already been covered in Unit B, Chapter 6 has many very short sections that can be combined and covered in almost any order.

UNIT D: Inference for Multiple Parameters

In Unit D we consider statistical inference for situations with multiple parameters: testing categorical variables with more than two categories (chi-square tests in Chapter 7), comparing means between more than two groups (ANOVA in Chapter 8), making inferences using the slope and intercept of a regression model (simple linear regression in Chapter 9), and building regression models with more than one explanatory variable (multiple regression in Chapter 10).

The Big Picture: Essential Synthesis

This section gives a quick overview of all of the units and asks students to put the pieces together with questions related to a case study on speed dating that draws on ideas from throughout the text.

Chapter P: Probability Basics

This is an optional chapter covering basic ideas of formal probability theory. The material in this chapter is independent of the other chapters and can be covered at any point in a course or omitted entirely.

Changes in the Third Edition

- **New Exercises.** The Third Edition includes over 200 completely new exercises, almost all of which use real data. As always, our goal has been to continue to try to find datasets and studies of interest to students and instructors.
- **Updated Exercises.** In addition to the many new exercises, this edition also includes over 100 exercises that have been updated with new data.
- **New Datasets.** We have added many new datasets and updated many others. There are now over 130 full datasets included in the materials (provided in many different software formats), as well as many additional smaller datasets discussed within the text.
- **Additional emphasis in Chapter 1.** We have added material in Chapter 1 to increase the emphasis on understanding the concepts of association, confounding, and causality.
- **Chapter 4 Reorganized.** Chapter 4 on hypothesis tests has been reorganized to focus more explicitly on the important step of locating the sample statistic in the randomization distribution. This allows us to place more emphasis on the connection between statistics that are extreme in the randomization distribution and low p-values, and then the connection to greater evidence against the null hypothesis.
- **Data Science.** Recognizing modern trends in data science, we have worked to include more large datasets with many cases and many variables. For example, the CollegeScores dataset has data on 37 variables for all 6141 post-secondary schools in the Department of Education's College Scorecard.
- **StatKey Enhanced.** We continue to add data sets and improve the functionality of the online interactive dynamic software StatKey.
- **And Much More!** We have also made additional edits to the text to improve the flow and clarity, keep it current, and respond to student and user feedback.

Tips for Students

- **Do the Exercises!** The key to learning statistics is to try lots of the exercises. We hope you find them interesting!
- **Videos** we have created video solutions for all examples and short video tutorials for all learning goals. These are available through WileyPLUS. If you need some help, check them out!
- **Partial Answers** Partial answers to the odd-numbered problems are included in the back of the book. These are *partial* answers, not full solutions or even complete answers. In particular, many exercises expect you to interpret or explain or show details, and you should do so! (Full solutions to the odd-numbered problems are included with WileyPLUS or the Student Solutions Manual.)

- **Exercises Referencing Exercises** Many of the datasets and studies included in this book are introduced and then referenced again later. When this happens, we include the earlier reference for your information, but *you should not need to refer back to the earlier reference.* All necessary information will be included in the later problem. The reference is there in case you get curious and want more information or the source citation.
- **Accuracy** Statistics is not an exact science. Just about everything is an approximation, with very few exactly correct values. Don't worry if your answer is slightly different from your friend's answer, or slightly different from the answer in the back of the book. Especially with the simulation methods of Chapters 3 and 4, a certain amount of variability in answers is a natural and inevitable part of the process.

Acknowledgments

The team at John Wiley & Sons, including Joanna Dingle, Tom Kulesa, Mike MacDougald, Laura Abrams, Katy Gabel, Maureen Shelburn, John LaVacca, Andrew Moore, Valerie Zaborski, Anne Scanlan-Rohrer, and Laurie Rosatone, has provided wonderful support and guidance, while also being great fun to work with. We especially appreciate the fact that they have shared our great enthusiasm for this project throughout our work together. We particularly thank Laurie Rosatone for her friendship and guidance throughout this project.

Ed Harcourt, Rich Sharp, Kevin Angstadt, and Yuxi Zhang are the programmers behind *StatKey*. We are incredibly appreciative of all of their efforts to bring our shared vision of these tools into a working reality and the many helpful suggestions for enhancements. Thanks also to John Lock for hosting the *lock5stat.com* website.

Ann Cannon did her usual fabulous job of the monumental task of accuracy checking the entire book, all exercise solutions, and making many helpful suggestions along the way.

Many people helped us collect lots of interesting datasets and applications that are so vital to a modern statistics text. These people include Edith Frazer, Zan Armstrong, Adam Pearce, Jim Vallandingham, Rick Cleary, Ivan Ramler, Tom DeRosa, Judy Graham, Bruce Frazer, Serge Onyper, Pamela Thatcher, Brad Baldwin, Laura Fonken, Jeremy Groves, Michael Frazer, Linda Casserly, Paul Doty, Ellen Langer, Gwen Cunningham, Noam Sobel, Lisa Torrey, and Dan Look. We appreciate that more and more researchers are willing to share their data to help students see the relevance of statistics to so many fields.

As we began this project, we benefitted greatly from many valuable discussions with colleagues and friends in the statistics education community who have been active in developing this new approach to teaching statistics, including Beth Chance, Laura Chihara, George Cobb, Bob del Mas, Michelle Everson, Joan Garfield, Jeff Hamrick, Tim Hesterberg, John Holcomb, Rebekah Isaak, Laura Le, Allan Rossman, Andrew Zieffler, and Laura Ziegler. We continue to be inspired by our many wonderful colleagues, new and old, in the statistics education community. Thanks also to Jeff Tecosky-Feldman and Dan Flath for their early support and encouragement for this project. Special thanks to Jessica Chapman for her excellent work on the Test Bank, Kathryn Dobeck for developing interactive lecture videos, Laura Le and Laura Ziegler for contributing additional activities, and Randall Pruim (and his students) for workiing on the Lock5withR packages and materials.

We thank Roxy Peck for the use of her home in Los Osos, CA for a sabbatical in the spring of 2011. Many of the words in this text were originally penned while appreciating her wonderful view of Morro Bay.

We thank our students who provided much valuable feedback on all drafts of the book, and who continue to help us improve the text and help us find interesting datasets.

We thank the many reviewers (listed at the end of this section) for their helpful comments, suggestions, and insights. They have helped make this text significantly better, and have helped give us confidence that this approach can work in a wide variety of settings.

We owe our love of both teaching and statistics at least in part to Ron Frazer, who was teaching innovative statistics classes as far back as the 60's and 70's. He read early versions of the book and was full of excitement and enthusiasm for the project. He spent 88 years enjoying life and laughing, and his constant smile and optimism are greatly missed.

The Second and Third Editions of this book were written amidst many wonderful additions to our growing family. We are especially grateful to Eugene Morgan, Amy Lock, and Nidhi Kohli for their love and support of this family and this project. All three share our love of statistics and have patiently put up with countless family conversations about the book. We are also very grateful for the love and joy brought into our lives by the next generation of Locks, all of whom have been born since the First Edition of this book: Axel, Cal, Daisy, and Zinnia Lock Morgan; Aadhya and Saumya Lock; and Jocelyn, Hailey, and Cody Lock, who we hope will also come to share our love of statistics!

Suggestions?

Our goal is to design materials that enable instructors to teach an excellent course and help students learn, enjoy, and appreciate the subject of statistics. If you have suggestions for ways in which we can improve the text and/or the available resources, making them more helpful, accurate, clear, or interesting, please let us know. We would love to hear from you! Our contact information is at *www.lock5stat.com*.

We hope you enjoy the journey!

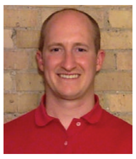
Robin H. Lock
Eric F. Lock
Kari Lock Morgan
Patti Frazer Lock
Dennis F. Lock

ABOUT THE AUTHORS

Robin H. Lock is Burry Professor of Statistics at St. Lawrence University. He is a Fellow of the American Statistical Association, past Chair of the Joint MAA-ASA Committee on Teaching Statistics, and a member of the committees that developed and later revised GAISE (Guidelines for Assessment and Instruction in Statistics Education). He has won the national Mu Sigma Rho Statistics Education award and was the inaugural recipient of the ASA's Waller Distinguished Teaching Career Award. He is also interested in applications of statistics in sports, especially ice hockey, where he can be found playing goalie for the SLU faculty/staff team.

Patti Frazer Lock is Cummings Professor of Mathematics at St. Lawrence University. She is a member of the Joint MAA-ASA Committee on Undergraduate Statistics Education, a member of the AMS-ASA-MAA-SIAM Joint Data Committee, and past Chair of the SIGMAA on Statistics Education. She received the 2016 MAA Seaway Section Award for Excellence in Teaching. She is a member of the Calculus Consortium and is a co-author on Hughes-Hallett's Calculus and Applied Calculus, Connally's Functions Modeling Change, and McCallum's Algebra and Multivariable Calculus texts. She is passionate about helping students succeed in, and enjoy, introductory courses in statistics and mathematics.

Kari Lock Morgan is Assistant Professor of Statistics at Pennsylvania State University. She received her Ph.D. in statistics from Harvard University in 2011, where she received the university-wide Derek C. Bok Teaching Award. She won the national 2018 Robert V. Hogg Award for Excellence in Teaching Introductory Statistics, gave a plenary address at the 2019 United States Conference on Teaching Statistics, and has served on the executive committee for the ASA Section on Statistics Education. She has a particular interest in causal inference, in addition to statistics education and simulation-based inference. She spent two years touring the world as a professional figure skater, and was an avid ultimate frisbee player. She now enjoys spending time with her husband, Eugene, and four kids, Axel, Cal, Daisy, and Zinnia.



Eric F. Lock is an Associate Professor of Biostatistics at the University of Minnesota School of Public Health. He received his Ph.D. in Statistics from the University of North Carolina in 2012, and spent two years doing a post doc in statistical genomics at Duke University. He has been an instructor for multiple introductory statistics courses and advanced graduate-level courses, and he has received a departmental teaching award. He has a particular interest in the analysis of high-dimensional data and Bayesian statistics, and is leading multiple projects on applications of statistics in molecular biology and medicine. He lives in Minneapolis with his wife, Nidhi Kohli, and two daughters Aadhya and Saumya.

Dennis F. Lock joined the Buffalo Bills as Director of Football Research and Strategy in 2019 after serving as Director of Football Analytics for the Miami Dolphins for the previous 5 years. He received his Ph.D. in statistics from Iowa State University where he served as an instructor and statistical consultant. While at Iowa State he helped design and implement a randomized study to compare the effectiveness of randomization and traditional approaches to teaching introductory statistics, and he received the Dan Mowrey Consulting Excellence Award. He and his wife, Amy, have three children: Jocelyn, Hailey, and Cody.

Reviewers for the Third Edition

Lynn A. Agre, MPH, PhD	*Rutgers University, Department of Statistics*	Amelia McNamara	*University of St. Thomas*
Karen Benway	*University of Vermont*	Andrew Poppick	*Carleton College*
Mary Daly	*Niagara University*	Whitney Zimmerman	*Penn State*
Amanda Mangum	*Converse College*		

Reviewers for the Second Edition

Alyssa Armstrong	*Wittenberg University*	Barb Barnet	*University of Wisconsin - Platteville*
Kevin Beard	*University of Vermont*	Barbara Bennie	*University of Wisconsin - La Crosse*
Karen Benway	*University of Vermont*	Dale Bowman	*University of Memphis*
Patricia Buchanan	*Penn State University*	Coskun Cetin	*California State University - Sacramento*
Jill A. Cochran	*Berry College*	Rafael Diaz	*California State University - Sacramento*
Kathryn Dobeck	*Lorain County Community College*	Brandi Falley	*Texas Women's University*
John Fieberg	*University of Minnesota*	Elizabeth Brondos Fry	*University of Minnesota*
Jennifer Galovich	*College of St. Benedict and St. John's University*	Steven T. Garren	*James Madison University*
Mohinder Grewal	*Memorial University of Newfoundland*	Robert Hauss	*Mt. Hood Community College*
James E. Helmreich	*Marist College*	Carla L. Hill	*Marist College*
Martin Jones	*College of Charleston*	Jeff Kollath	*Oregon State University*
Paul Kvam	*University of Richmond*	David Laffie	*(formerly) California State University - East Bay*
Bernadette Lanciaux	*Rochester Institute of Technology*	Anne Marie S. Marshall	*Berry College*
Gregory Mathews	*Loyola University Chicago*	Scott Maxwell	*University of Notre Dame*
Kathleen McLaughlin	*University of Connecticut*	Sarah A Mustillo	*University of Notre Dame*
Elaine T. Newman	*Sonoma State University*	Rachel Rader	*Ohio Northern University*
David M. Reineke	*University of Wisconsin - La Crosse*	Rachel Roe-Dale	*Skidmore College*
Kimberly A. Roth	*Juniata College*	Dinesh Sharma	*James Madison University*
Alla Sikorskii	*Michigan State University*	Karen Starin	*Columbus State Community College*
Paul Stephenson	*Grand Valley State University*	Asokan Variyath	*Memorial University*
Lissa J. Yogan	*Valparaiso University*	Laura Ziegler	*Iowa State University*

Reviewers and Class Testers for the First Edition

Wendy Ahrensen	*South Dakota St.*	Diane Benner	*Harrisburg Comm. College*
Steven Bogart	*Shoreline Comm. College*	Mark Bulmer	*University of Queensland*
Ken Butler	*Toronto-Scarborough*	Ann Cannon	*Cornell College*
John “Butch” Chapelle	*Brookstone School*	George Cobb	*Mt. Holyoke University*
Steven Condly	*U.S. Military Academy*	Salil Das	*Prince Georges Comm. College*
Jackie Dietz	*Meredith College*	Carolyn Dobler	*Gustavus Adolphus*
Robert Dobrow	*Carleton College*	Christiana Drake	*Univ. of Calif. - Davis*
Katherine Earles	*Wichita State*	Laura Estersohn	*Scarsdale High School*
Karen A. Estes	*St. Petersburg College*	Soheila Fardanesh	*Towson University*
Diane Fisher	*Louisiana - Lafayette*	Brad Fulton	*Duke University*
Steven Garren	*James Madison Univ.*	Mark Glickman	*Boston University*
Brenda Gunderson	*Univ. of Michigan*	Aaron Gutknecht	*Tarrant County C.C.*
Ian Harris	*Southern Methodist Univ.*	John Holliday	*North Georgia College*
Pat Humphrey	*Georgia Southern Univ.*	Robert Huotari	*Glendale Comm.College*
Debra Hydorn	*Univ. of Mary Washington*	Kelly Jackson	*Camden County College*
Brian Jersky	*Macquarie University*	Matthew Jones	*Austin Peay St. Univ.*
James Lang	*Valencia Comm. College*	Lisa Lendway	*University of Minnesota*
Stephen Lee	*University of Idaho*	Christopher Malone	*Winona State*
Catherine Matos	*Clayton State*	Billie May	*Clayton State*

Monnie McGee	*Southern Methodist Univ.*
Matthew Mitchell	*Florida St.-Jacksonville*
Perpetua Lynne Nielsen	*Brigham Young University*
Nabendu Pal	*Louisiana - Lafayette*
Iwan Praton	*Franklin & Marshall College*
Christian Rau	*Monash University*
Thomas Roe	*South Dakota State*
Yuliya Romanyuk	*King's University College*
Edith Seier	*East Tennessee State*
Benjamin Sherwood	*University of Minnesota*
Dalene Stangl	*Duke University*
Sheila Weaver	*Univ. of Vermont*
Alison Weir	*Toronto-Mississauaga*
Rebecca Wong	*West Valley College*
William Meisel	*Florida St.-Jacksonville*
Lori Murray	*Western University*
Julia Norton	*Cal. State - East Bay*
Alison Paradise	*Univ. of Puget Sound*
Guoqi Qian	*Univ. of Melbourne*
Jerome Reiter	*Duke University*
Rachel Roe-Dale	*Skidmore College*
Charles Scheim	*Hartwick College*
Therese Shelton	*Southwestern Univ.*
Sean Simpson	*Westchester Comm. College*
Robert Stephenson	*Iowa State*
John Weber	*Georgia Perimeter College*
Ian Weir	*Univ. of the West of England*
Laura Ziegler	*University of Minnesota*

CONTENTS

UNIT A

Data

"For Today's Graduate, Just One Word: Statistics"

New York Times headline, August 5, 2009

UNIT OUTLINE

1 Collecting Data

2 Describing Data

Essential Synthesis

In this unit, we learn how to collect and describe data. We explore how data collection influences the types of conclusions that can be drawn, and discover ways to summarize and visualize data.

iStock.com/petesaloutos

iStock.com/KeithSzafranski

Al Diaz/Miami Herald/Getty Images

CHAPTER 1

Collecting Data

“You can’t fix by analysis what you bungled by design.”

Richard Light, Judith Singer, and John Willett in *By Design*

CHAPTER OUTLINE

Questions and Issues

Here are some of the questions and issues we will discuss in this chapter:

- Is there a "sprinting gene"?
- Does tagging penguins for identification purposes harm them?
- Do humans subconsciously give off chemical signals (pheromones)?
- What proportion of people using a public restroom wash their hands?
- If parents could turn back time, would they still choose to have children?
- Why do adolescent spiders engage in foreplay?
- How broadly do experiences of parents affect their future children?
- What percent of college professors consider themselves "above-average" teachers?
- Does giving lots of high fives to teammates help sports teams win?
- Which is better for peak performance: a short mild warm-up or a long intense warm-up?
- Are city dwellers more likely than country dwellers to have mood and anxiety disorders?
- Is there truth to the saying "beauty sleep"?
- What percent of young adults in the US move back in with their parents?
- Does turning up the music in a bar cause people to drink more beer?
- Is your nose getting bigger?
- Does watching cat videos improve mood?
- Does sleep deprivation hurt one's ability to interpret facial expressions?
- Do artificial sweeteners cause weight gain?
- Does late night eating impair concentration?
- Does eating organic food improve your health?

1.1 THE STRUCTURE OF DATA

We are being inundated with data. It is estimated that the amount of new technical information is doubling every two years, and that over 7.2 zettabytes (that's 7.2×10^{21} bytes) of unique new information is being generated every year.[1] That is more than was generated during the entire 5000-year period before you were born. An incredible amount of data is readily available to us on the Internet and elsewhere. The people who are able to analyze this information are going to have great jobs and are going to be valuable in virtually every field. One of the wonderful things about statistics is that it is relevant in so many areas. Whatever your focus and your future career plans, it is likely that you will need statistical knowledge to make smart decisions in your field and in everyday life. As we will see in this text, effective collection and analysis of data can lead to very powerful results.

***Statistics** is the science of collecting, describing, and analyzing data.* In this chapter, we discuss effective ways to *collect* data. In Chapter 2, we discuss methods to *describe* data. The rest of the chapters are devoted to ways of *analyzing* data to make effective conclusions and to uncover hidden phenomena.

DATA 1.1

A Student Survey

For several years, a first-day survey has been administered to students in an introductory statistics class at one university. Some of the data for a few of the students are displayed in Table 1.1. A more complete table with data for 362 students and 17 variables can be found in the file **StudentSurvey**.[2] ■

Cases and Variables

The subjects/objects that we obtain information about are called the *cases* or *units* in a dataset. In the **StudentSurvey** dataset, the cases are the students who completed the survey. Each row of the dataset corresponds to a different case.

A *variable* is any characteristic that is recorded for each case. Each column of our dataset corresponds to a different variable. The data in Table 1.1 show eight variables (in addition to the ID column), each describing a different characteristic of the students taking the survey.

Table 1.1 ***Partial results from a student survey***

ID	Sex	Smoke	Award	Exercise	TV	GPA	Pulse	Birth
1	M	No	Olympic	10	1	3.13	54	4
2	F	Yes	Academy	4	7	2.5	66	2
3	M	No	Nobel	14	5	2.55	130	1
4	M	No	Nobel	3	1	3.1	78	1
5	F	No	Nobel	3	3	2.7	40	1
6	F	No	Nobel	5	4	3.2	80	2
7	F	No	Olympic	10	10	2.77	94	1
8	M	No	Olympic	13	8	3.3	77	1
9	F	No	Nobel	3	6	2.8	60	2
10	F	No	Nobel	12	1	3.7	94	8

[1] "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," *https://www.emc.com/leadership/digital-universe/2012iview/index.htm*. Accessed June 2020.

[2] Most datasets used in this text, and descriptions, are available electronically. They can be found at *www.wiley.com/college/lock* and at *www.lock5stat.com*. See the Preface for more information. Descriptions of many datasets can also be found in Appendix B.

Cases and Variables

We obtain information about **cases** or **units** in a dataset, and generally record the information for each case in a row of a data table.

A **variable** is any characteristic that is recorded for each case. The variables generally correspond to the columns in a data table.

In any dataset, it is important to understand exactly what each variable is measuring and how the values are coded. For the data in Table 1.1, the first column is ID, to provide an identifier for each of the individuals in the study. In addition, we have:

Sex	M for male and F for female[3]
Smoke	Does the student smoke: yes or no
Award	Award the student prefers to win: Academy Award, Olympic gold medal, or Nobel Prize
Exercise	Number of hours spent exercising per week
TV	Number of hours spent watching television per week
GPA	Current grade point average on a 4-point scale
Pulse	Pulse rate in number of beats per minute at the time of the survey
Birth	Birth order: 1 for first/oldest, 2 for second born, etc.

Example 1.1

Explain what each variable tells us about the student with ID 1 in the first row of Table 1.1.

Solution

Student 1 is a male who does not smoke and who would prefer to win an Olympic gold medal over an Academy Award or a Nobel Prize. He says that he exercises 10 hours a week, watches television one hour a week, and that his grade point average is 3.13. His pulse rate was 54 beats per minute at the time of the survey, and he is the fourth oldest child in his family.

Categorical and Quantitative Variables

In determining the most appropriate ways to summarize or analyze data, it is useful to classify variables as either *categorical* or *quantitative*.

Categorical and Quantitative Variables

A **categorical variable** divides the cases into groups, placing each case into exactly one of two or more categories.

A **quantitative variable** measures or records a numerical quantity for each case. Numerical operations like adding and averaging make sense for quantitative variables.

[3] We acknowledge that this binary dichotomization is not a complete or inclusive representation of reality. However, the binary option is frequently used in data collection for ease of analysis and for confidentiality purposes. We apologize for the lack of inclusivity in situations involving sex and gender in this text.

We may use numbers to code the categories of a categorical variable, but this does not make the variable quantitative unless the numbers have a quantitative meaning. For example, "sex" is categorical even if we choose to record the results as 1 for male and 2 for female, since we are more likely to be interested in how many are in each category rather than an average numerical value. In other situations, we might choose to convert a quantitative variable into categorical groups. For example, "household income" is quantitative if we record the specific values but is categorical if we instead record only an income category ("low," "medium," "high") for each household.

Example 1.2

Classify each of the variables in the student survey data in Table 1.1 as either categorical or quantitative.

Solution

Note that the ID column is neither a quantitative nor a categorical variable. A dataset often has a column with names or ID numbers that are for reference only.

- *Sex* is categorical since it classifies students into categories of male and female.
- *Smoke* is categorical since it classifies students as smokers or nonsmokers.
- *Award* is categorical since students are classified depending on which award is preferred.
- *Exercise*, *TV*, *GPA*, and *Pulse* are all quantitative since each measures a numerical characteristic of each student. It makes sense to compute an average for each variable, such as an average number of hours of exercise a week.
- *Birth* is a somewhat ambiguous variable, as it could be considered either quantitative or categorical depending on how we use it. If we want to find an average birth order, we consider the variable quantitative. However, if we are more interested in knowing how many first-borns, how many second-borns, and so on, are in the data, we consider the variable categorical. Either answer is acceptable.

Investigating Variables and Relationships between Variables

In this book, we discuss ways to describe and analyze a single variable and to describe and analyze relationships between two or more variables. For example, in the student survey data, we might be interested in the following questions, each about a single variable:

- What percentage of students smoke?
- What is the average number of hours a week spent exercising?
- Are there students with unusually high or low pulse rates?
- Which award is the most desired?
- How does the average GPA of students in the survey compare to the average GPA of all students at this university?

Often the most interesting questions arise as we look at relationships between variables. In the student survey data, for example, we might ask the following questions about relationships between variables:

- Who smokes more, males or females?
- Do students who exercise more tend to prefer an Olympic gold medal? Do students who watch lots of television tend to prefer an Academy Award?

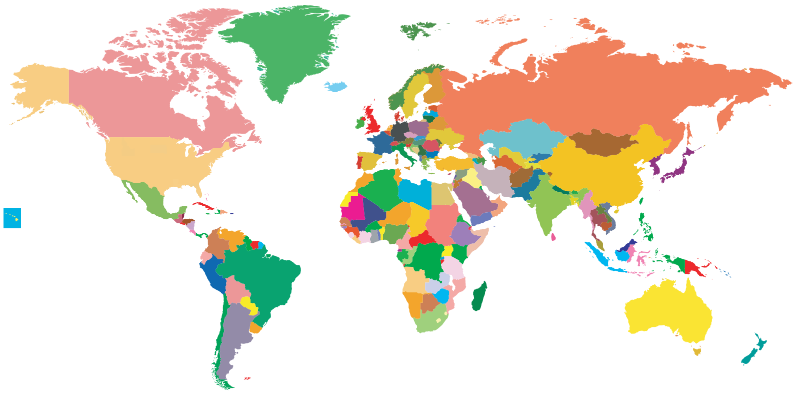
- Do males or females watch more television?
- Do students who exercise more tend to have lower pulse rates?
- Do first-borns generally have higher grade point averages?

These examples show that relationships might be between two categorical variables, two quantitative variables, or a quantitative and a categorical variable. In the following chapters, we examine statistical techniques for exploring the nature of relationships in each of these situations.

DATA 1.2

Data on Countries

As of this writing, there are 217 countries listed by the World Bank.[4] A great deal of information about these countries (such as energy use, birth rate, life expectancy) is in the full dataset under the name **AllCountries**. ■



redmal/Getty Images

Countries of the world

Example 1.3

The dataset **AllCountries** includes information on the percent of people in each country with access to the Internet.

(a) Data from the Principality of Andorra were used to determine that 98.9% of Andorrans have access to the Internet, the highest rate of any country. What are the cases in the data from Andorra? What variable is used? Is it categorical or quantitative?

(b) In the **AllCountries** dataset, we record the percent of people with access to the Internet for each country. What are the cases in that dataset? What is the relevant variable? Is it categorical or quantitative?

Solution

(a) For determining the rate of Internet usage in Andorra, the cases are people in Andorra, and the relevant variable is whether or not each person has access to the Internet. This is a categorical variable.

(b) In the **AllCountries** dataset, the cases are the countries of the world. The variable is the proportion with access to the Internet. For each country, we record a numerical value. These values range from a low of 1.3% in Eritrea to the high of 98.9% in Andorra, and the average is 54.47%. This is a quantitative variable.

[4] *http://data.worldbank.org/*. Data include information on both countries and economies, accessed June 2019.

As we see in the previous example, we need to think carefully about what the cases are and what is being recorded in each case in order to determine whether a variable is categorical or quantitative.

Example 1.4

In later chapters, we examine some of the following issues using the data in **AllCountries**. Indicate whether each question is about a single variable or a relationship between variables. Also indicate whether the variables are quantitative or categorical.

(a) How much energy does the average country use in a year?

(b) Do countries larger in area tend to have a more rural population?

(c) What is the relationship, if any, between a country's government spending on the military and on health care?

(d) Is the birth rate higher in developed or undeveloped countries?

(e) Which country has the highest percent of elderly people?

Solution

(a) The amount of energy used is a single quantitative variable.

(b) Both size and percent rural are quantitative variables, so this is a question about a relationship between two quantitative variables.

(c) Spending on the military and spending on health care are both quantitative, so this is another question about the relationship between two quantitative variables.

(d) Birth rate is a quantitative variable and whether or not a country is developed is a categorical variable, so this is asking about a relationship between a quantitative variable and a categorical variable.

(e) Because the cases are countries, percent elderly is a single quantitative variable.

Using Data to Answer a Question

The **StudentSurvey** and **AllCountries** datasets contain lots of information and we can use that information to learn more about students and countries. Increasingly, in this data-driven world, we have large amounts of data and we want to "mine" it for valuable information. Often, however, the order is reversed: We have a question of interest and we need to collect data that will help us answer that question.

iStock.com/petesaloutos

Is there a "sprinting gene"?

Example 1.5

Is There a "Sprinting Gene"?

A gene called *ACTN3* encodes a protein which functions in fast-twitch muscles. Some people have a variant of this gene that cannot yield this protein. (So we might call the gene variant a possible *non*-sprinting gene.) To address the question of whether this gene is associated with sprinting ability, geneticists tested people from three different groups: world-class sprinters, world-class marathon runners, and a control group of non-athletes. In the samples tested, 6% of the sprinters had the gene variant, compared with 18% of the non-athletes and 24% of the marathon runners. This study[5] suggests that sprinters are less likely than non-sprinters to have the gene variant.

(a) What are the cases and variables in this study? Indicate whether each variable is categorical or quantitative.

(b) What might a table of data look like for this study? Give a table with a possible first two cases filled in.

Solution

(a) The cases are the people included in the study. One variable is whether the individual has the gene variant or not. Since we record simply "yes" or "no," this is a categorical variable. The second variable keeps track of the group to which the individual belongs. This is also a categorical variable, with three possible categories (sprinter, marathon runner, or non-athlete). We are interested in the relationship between these two categorical variables.

(b) The table of data must record answers for each of these variables and may or may not have an identifier column. Table 1.2 shows a possible first two rows for this dataset.

Table 1.2 ***Possible table to investigate whether there is a sprinter's gene***

Name	Gene Variant	Group
Allan	Yes	Marathon runner
Beth	No	Sprinter
. . .	. . .	. . .

Example 1.6

What's a Habanero?

A habanero chili is an extremely spicy pepper (roughly 500 times hotter than a jalapeño) that is used to create fiery food. The vice president of product development and marketing for the Carl's Jr. restaurant chain[6] is considering adding a habanero burger to the menu. In developing an advertising campaign, one of the issues he must deal with is whether people even know what the term "habanero" means. He identifies three specific questions of interest and plans to survey customers who visit the chain's restaurants in various parts of the country.

- What proportion of customers know and understand what "habanero" means?
- What proportion of customers are interested in trying a habanero burger?
- How do these proportions change for different regions of the country?

[5] Yang, N., et al., "ACTN3 genotype is associated with human elite athletic performance," *American Journal of Human Genetics*, September 2003; 73: 627–631.

[6] With thanks to Bruce Frazer.

(a) Identify the cases for the data he collects.

(b) Describe three variables that should be included in the dataset.

Solution

(a) The cases in the habanero marketing study are the individual customers that respond to the survey request.

(b) Here are three variables that should be included in the data:

- *Know* = yes or no, depending on whether the customer knows the term "habanero"
- *Try* = yes or no, to indicate the customer's willingness to try a habanero burger
- *Region* = area in the country where the customer lives

All three variables are categorical.

Notice that for each case (customer), we record a value for each of the variables. Each customer would be represented in the dataset by a different row in the data table.

iStock.com/KeithSzafranski

Does tagging penguins harm them?

DATA 1.3

Tagging Penguins

Do metal tags on penguins harm them? Scientists trying to tell penguins apart have several ways to tag the birds. One method involves wrapping metal strips with ID numbers around the penguin's flipper, while another involves electronic tags. Neither tag seems to physically harm the penguins. However, since tagged penguins are used to study *all* penguins, scientists wanted to determine whether the metal tags have any significant effect on the penguins. Data were collected over a 10-year time span from a sample of 100 penguins that were randomly given either metal or electronic tags. This included information on number of chicks, survival over the decade, and length of time on foraging trips.[7] ■

Example 1.7

In the study on penguin tags:

(a) What are the cases? What are the variables? Identify each variable as categorical or quantitative.

[7] Saraux, C., et al., "Reliability of flipper-banded penguins as indicators of climate change," *Nature*, January 2011; 469: 203–206.

(b) What information do the scientists hope to gain from the data? Describe at least one question in which they might be interested.

Solution

(a) The cases are the tagged penguins. The variables are the type of tag (categorical), number of chicks (quantitative), survival or not (categorical), and length of time on foraging trips (quantitative).

(b) The scientists want to determine whether there is a relationship between the type of tag and any of the other variables. For example, they might be interested in knowing whether survival rates differ between the penguins with the metal tags and penguins with the electronic tags.

In Example 1.7, we are investigating whether one of the variables (the type of tag) helps us explain or predict values of the other variables. In this situation, we call the type of tag the *explanatory variable* and the other variables the *response variables*. One way to remember these names is the explanatory variable helps explain the response variable, and the response variable responds to the explanatory variable.

Explanatory and Response Variables

If we are using one variable to help us understand or predict values of another variable, we call the former the **explanatory variable** and the latter the **response variable**.

Example 1.8

In Example 1.4, we considered the following three questions about relationships between variables in the **AllCountries** dataset. Identify the explanatory variable and the response variable if it makes sense to do so.

(a) Do countries larger in area tend to have a more rural population?

(b) Is the birth rate higher in developed or undeveloped countries?

(c) What is the relationship, if any, between a country's government spending on the military and on health care?

Solution

(a) The question indicates that we think area might influence the percent of a country that is rural, so we call area the explanatory variable and percent rural the response variable.

(b) The question indicates that we think whether or not a country is developed might influence the birth rate, so the explanatory variable is whether the country is developed or undeveloped and the response variable is birth rate.

(c) There is no indication in this situation of why we might identify either of the two variables (spending on military and spending on health care) as explanatory or as response. In a relationship between two variables, we don't always identify one as the explanatory variable and one as the response variable.

Different Ways to Answer a Broad Question

A pheromone is a chemical signal given off by one member of a species that influences other members of the species. Many studies (involving data, of course!) provide evidence that different types of animals give off pheromones. It is currently under debate whether humans also communicate subconsciously through pheromones. How might we collect data to answer the question of whether there are human pheromones? We might start by narrowing the question to one that is not so general. For example, are there pheromones in female tears that affect the behavior of males?

Several studies[8] suggest that the scent of tears from a crying woman may reduce sexual interest in men. However, to determine whether this effect is caused subconsciously by pheromones (rather than by obvious social influences), we need to think carefully about how to collect data. How might you collect data to answer this question? Three different methods were used in the studies. See what you think of the three methods.

- In one study, 25 men in their twenties had a pad attached to their upper lip that contained either tears collected from women who watched sad films or a salt solution that had been trickled down the same women's faces. Neither substance had a perceptible odor. The men who had tears on the upper lip rated female faces as less sexually alluring than the men who had salt solution on the upper lip.
- In a second study, 50 men who sniffed women's tears showed reduced levels of testosterone relative to levels after sniffing a salt solution.
- In a third study involving 16 men, those who sniffed female tears displayed significantly reduced brain-cell activity in areas that had reacted strongly to an erotic movie, whereas those who sniffed a salt solution did not show the same reduced activity.

Example 1.9

For each of the three studies on women's tears, state the explanatory and response variables.

Solution

In all three studies, the explanatory variable is whether tears or a salt solution is used. In the first study, the response variable is how sexually alluring males rated female faces, in the second study it is testosterone levels, and in the third study it is brain cell activity.

All three of these studies describe data collected in a careful way to answer a question. How to collect data in a way that helps us understand real phenomena is the subject of the rest of this chapter.

We have described several datasets, studies, and questions in this section, involving students, countries, sprinter genes, habanero burgers, penguins, and pheromones. If you are intrigued by any of these questions, keep reading! We examine all of them in more detail in the pages ahead.

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- Recognize that a dataset consists of cases and variables
- Identify variables as either categorical or quantitative
- Determine explanatory and response variables where appropriate
- Describe how data might be used to address a specific question
- Recognize that understanding statistics allows us to investigate a wide variety of interesting phenomena

[8] Gelstein, S., et al., "Human Tears Contain a Chemosignal," *Science*, January 2011; 331(6014): 226–230.

Exercises for Section 1.1

SKILL BUILDER 1

For the situations described in Exercises 1.1 to 1.6:

(a) What are the cases?

(b) What is the variable and is it quantitative or categorical?

1.1 People in a city are asked if they support a new recycling law.

1.2 Record the percentage change in the price of a stock for 100 stocks publicly traded on Wall Street.

1.3 Collect data from a sample of teenagers with a question that asks "Do you eat at least five servings a day of fruits and vegetables?"

1.4 Measure the shelf life of bunches of bananas (the number of days until the bananas go bad) for a large sample.

1.5 Estimate the bending strength of beams by bending 10 beams until they break and recording the force at which the beams broke.

1.6 Record whether or not the literacy rate is over 75% for each country in the world.

SKILL BUILDER 2

In Exercises 1.7 to 1.10, a relationship between two variables is described. In each case, we can think of one variable as helping to explain the other. Identify the explanatory variable and the response variable.

1.7 Lung capacity and number of years smoking cigarettes

1.8 Amount of fertilizer used and the yield of a crop

1.9 Blood alcohol content (BAC) and number of alcoholic drinks consumed

1.10 Year and the world record time in a marathon

1.11 Student Survey Variables Data 1.1 introduced the dataset **StudentSurvey**, and Example 1.2 identified seven of the variables in that dataset as categorical or quantitative. The remaining variables are:

Year	FirstYear, Sophomore, Junior, Senior
Height	In inches
Weight	In pounds
Siblings	Number of siblings the person has
VerbalSAT	Score on the Verbal section of the SAT exam
MathSAT	Score on the Math section of the SAT exam
SAT	Sum of the scores on the Verbal and Math sections of the SAT exam
HigherSAT	Which is higher, Math SAT score or Verbal SAT score?

(a) Indicate whether each variable is quantitative or categorical.

(b) List at least two questions we might ask about any one of these individual variables.

(c) List at least two questions we might ask about relationships between any two (or more) of these variables.

1.12 Countries of the World Information about the world's countries is given in **AllCountries**, introduced in Data 1.2 on page 7. You can find a description of the variables in Appendix B. For the full dataset:

(a) Indicate which of the variables are quantitative and which are categorical.

(b) List at least two questions we might ask about any one of these individual variables.

(c) List at least two questions we might ask about relationships between any two (or more) of these variables.

1.13 Goldilocks Effect: Read to Your Kids! The American Academy of Pediatrics recommends that parents begin reading to their children soon after birth, and that parents set limits on screen time. A new study[9] reinforces these recommendations. In the study, 27 four-year-olds were presented with stories in three different formats: audio (sound only), illustrated (sound and pictures), and animated (sound and animation). During the presentations, a magnetic resonance imaging (MRI) machine measured each child's brain connectivity. The researchers found a "Goldilocks effect," in which audio was too cold (with low brain connectivity as the children strained to understand) and animation was too hot (with low brain connectivity as the animation did all the work for the children). The highest connectivity (just right!) was found with the illustrated format, which simulates reading a book to a child.

[9] Hutton J et al., "Differences in functional brain network connectivity during stories presented in audio, illustrated, and animated format in preschool-age children," *Brain Imaging and Behavior*, October 30, 2018.

(a) What is the explanatory variable? Is it categorical or quantitative?

(b) What is the response variable? Is it categorical or quantitative?

(c) How many cases are there?

1.14 Female Gamers Face Sexual Harassment A research firm[10] questioned 1151 female gamers in Great Britain and found that 40% had received obscene messages while playing online. In addition to asking whether they had received obscene messages, the gamers were also asked how many hours a week they played, and whether they felt there were enough strong female characters in games.

(a) What are the cases in this study?

(b) What are the variables? Indicate whether each is categorical or quantitative.

(c) How many rows and how many columns will the dataset have if cases are rows and variables are columns?

1.15 Active Learning vs Passive Learning: Which is Best? Active learning in a classroom implies that students are actively involved and working during class time (either individually, in pairs, or in groups) while passive learning indicates that students are primarily taking notes while the instructor lectures. A recent study[11] measured students actual learning under these two formats as well as their feelings of learning. The study was very well designed: students in a college physics course were randomly assigned to a class period with either active learning or passive learning, the same content and handouts were used in both, and both instructors were highly rated. After the class, students were asked to rate how much they thought they had learned (on a 5-point Likert scale) and they also took a 20-question multiple choice exam to test how much they had actually learned. The results were very interesting: students *thought* that they learned more in the passive learning class but they *actually* learned more in the active learning class.

(a) What are the cases?

(b) What are the variables? Indicate whether each variable is quantitative or categorical.

(c) Indicate explanatory and response variables.

(d) There were 154 students in the passive learning lecture and 142 students in the active learning class. Indicate how many rows and how many columns the dataset will have if cases are rows and variables are columns.

1.16 Spider Sex Play Spiders regularly engage in spider foreplay that does not culminate in mating. Male spiders mature faster than female spiders and often practice the mating routine on not-yet-mature females. Since male spiders run the risk of getting eaten by female spiders, biologists wondered why spiders engage in this behavior. In one study,[12] some spiders were allowed to participate in these near-matings, while other maturing spiders were isolated. When the spiders were fully mature, the scientists observed real matings. They discovered that if either partner had participated at least once in mock sex, the pair reached the point of real mating significantly faster than inexperienced spiders did. (Mating faster is, apparently, a real advantage in the spider world.) Describe the variables, indicate whether each variable is quantitative or categorical, and indicate the explanatory and response variables.

1.17 Hormones and Fish Fertility When women take birth control pills, some of the hormones found in the pills eventually make their way into lakes and waterways. In one study, a water sample was taken from various lakes. The data indicate that as the concentration of estrogen in the lake water goes up, the fertility level of fish in the lake goes down. The estrogen level is measured in parts per trillion (ppt) and the fertility level is recorded as the percent of eggs fertilized. What are the cases in this study? What are the variables? Classify each variable as either categorical or quantitative.

1.18 Fast-Twitch Muscles and Race Example 1.5 studied a variant of the gene *ACTN3* which inhibits fast-twitch muscles and seems to be less prevalent in sprinters. A separate study[13] indicated ethnic differences: Approximately 20% of a sample of Caucasians, approximately 25% of a sample of Asians, and approximately 1% of a sample of Africans had the gene variant. What are the variables in this study? Classify each as categorical or quantitative.

[10] "Research: One in 3 Female Gamers Face Gender Discrimination, 32% Deal with Sexual Harassment," *Bryter-research.co.uk*, October 2019.

[11] Deslauriers L, et al., "Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom," *PNAS*, 116(39), September 24, 2019.

[12] Pruitt, J., paper presented at the Society for Integrative and Comparative Biology Annual Meeting, January 2011, and reported in "For spiders, sex play has its pluses," *Science News*, January 29, 2011.

[13] North, K., et al., "A common nonsense mutation results in α-actinin-3 deficiency in the general population," *Nature Genetics*, April 1999; 21(4): 353–354.

1.19 **Largest Cities in the World** Seven of the ten largest cities in the world are in the Eastern Hemisphere (including the largest: Tokyo, Japan) and three are in the Western Hemisphere.[14] Table 1.3 shows the populations, in millions of people, for these cities.

(a) How many cases are there in this dataset? How many variables are there and what are they? Is each categorical or quantitative?

(b) Display the information in Table 1.3 as a dataset with cases as rows and variables as columns.

Table 1.3 ***Population, in millions, of the world's largest cities***

Eastern hemisphere:	37, 26, 23, 22, 21, 21, 21
Western hemisphere:	21, 20, 19

1.20 **How Fast Do Homing Pigeons Go?** Homing pigeons have an amazing ability to find their way home over extremely long distances. How fast do they go on these trips? In the 2019 Midwest Classic, held in Topeka, Kansas, the fastest bird went 1676 YPM (yards per minute), which is about 56 miles per hour.[15] The top seven finishers included three female pigeons (Hens) and four male pigeons (Cocks). Their speeds, in YPM, are given in Table 1.4.

(a) How many cases are there in this dataset? How many variables are there and what are they? Is each variable categorical or quantitative?

(b) Display the information as a dataset with cases as rows and variables as columns.

Table 1.4 ***Speed of homing pigeons, in yards per minute***

Hens:	1676, 1452, 1449
Cocks:	1458, 1435, 1418, 1413

1.21 **Pigeon Racing** Exercise 1.20 gives the speed of the top seven finishers in the 2019 Midwest Classic homing pigeon race. In fact, 1412 pigeons finished the race, and their home loft, sex, distance, and speed were all recorded. A loft may have several different pigeons finish the race.

(a) How many cases are in this dataset?

(b) How many variables are there? How many of these variables are categorical? How many are quantitative?

(c) How many rows and how many columns will the dataset have?

1.22 **Trans-Generational Effects of Diet** Can experiences of parents affect future children? New studies[16] suggest that they can: Early life experiences of parents appear to cause permanent changes in sperm and eggs. In one study, some male rats were fed a high-fat diet with 43% of calories from fat (a typical American diet), while others were fed a normal healthy rat diet. Not surprisingly, the rats fed the high-fat diet were far more likely than the normal-diet rats to develop metabolic syndrome (characterized by such things as excess weight, excess fat, insulin resistance, and glucose intolerance.) What surprised the scientists was that the daughters of these rats were also far more likely to develop metabolic syndrome than the daughters of rats fed healthy diets. None of the daughters and none of the mothers ate a high-fat diet and the fathers did not have any contact with the daughters. The high-fat diet of the fathers appeared to cause negative effects for their daughters. What are the two main variables in this study? Is each categorical or quantitative? Identify the explanatory and response variables.

1.23 **Trans-Generational Effects of Environment** In Exercise 1.22, we ask whether experiences of parents can affect future children, and describe a study that suggests the answer is yes. A second study, described in the same reference, shows similar effects. Young female mice were assigned to either live for two weeks in an enriched environment or not. Matching what has been seen in other similar experiments, the adult mice who had been exposed to an enriched environment were smarter (in the sense that they learned how to navigate mazes faster) than the mice that did not have that experience. The other interesting result, however, was that the offspring of the mice exposed to the enriched environment were also smarter than the offspring of the other mice, even though none of the offspring were exposed to an enriched environment themselves. What are the two main variables in this study? Is each categorical or quantitative? Identify explanatory and response variables.

[14] *http://www.worldatlas.com/citypops.htm*. Accessed June 2015.

[15] Data downloaded from the Midwest Homing Pigeon Association final race report at *http://www.midwesthpa.com/MIDFinalReports.htm*.

[16] Begley, S., "Sins of the Grandfathers," *Newsweek*, November 8, 2010; 48–50.

1.24 Pennsylvania High School Seniors The data in **PASeniors** shows results for a sample of 457 high school seniors in the state of Pennsylvania, selected at random from all students who participated in the Census at Schools project[17] between 2010 and 2019. Each of the questions below relate to information in this dataset. Determine whether the answer to each question gives a value for a quantitative variable, a categorical variable, or is not a value for a variable for this dataset.

(a) What mode of transportation do you use to get to school?

(b) Do you have any allergies?

(c) What proportion of students in this sample are vegetarians?

(d) How many hours did you spend last week working at a paid job?

(e) What is the difference between typical hours of sleep you get on school nights and non-school nights?

(f) What is the maximum time (in minutes) that a student in this sample needs to get to school?

(g) If you could have a super power would you choose invisibility, telepathy, super strength, ability to fly, or ability to freeze time?

1.25 US College Scorecard The data in **CollegeScores** contains information from the US Department of Education's College Scorecard[18] on all postsecondary educational institutions in the US. Each of the questions below relate to information in this dataset. Determine whether the answer to each question gives a value for a quantitative variable, a categorical variable, or is not a value for a variable for this dataset.

(a) What is the total tuition and fees for in-state students at the school?

(b) How many of these schools are located in the Northeast?

(c) Is the school public, private, or for profit?

(d) How many undergraduates are enrolled at the school?

(e) What percentage of undergraduates at the school are part-time students?

(f) Which school has the highest average faculty salary?

1.26 Hookahs and Health Hookahs are waterpipes used for smoking flavored tobacco. One study[19] of 3770 university students in North Carolina found that 40% had smoked a hookah at least once, with many claiming that the hookah smoke is safer than cigarette smoke. However, a second study observed people at a hookah bar and recorded the length of the session, the frequency of puffing, and the depth of inhalation. An average session lasted one hour and the smoke inhaled from an average session was equal to the smoke in more than 100 cigarettes. Finally, a third study measured the amount of tar, nicotine, and heavy metals in samples of hookah smoke, finding that the water in a hookah filters out only a very small percentage of these chemicals. Based on these studies and others, many states are introducing laws to ban or limit hookah bars. In each of the three studies, identify the individual cases, the variables, and whether each variable is quantitative or categorical.

1.27 Is Your Nose Getting Bigger? Next time you see an elderly man, check out his nose and ears! While most parts of the human body stop growing as we reach adulthood, studies show that noses and ears continue to grow larger throughout our lifetime. In one study[20] examining noses, researchers report "Age significantly influenced all analyzed measurements:" including volume, surface area, height, and width of noses. The sex of the 859 participants in the study was also recorded, and the study reports that "male increments in nasal dimensions were larger than female ones."

(a) How many variables are mentioned in this description?

(b) How many of the variables are categorical? How many are quantitative?

(c) If we create a dataset of the information with cases as rows and variables as columns, how many rows and how many columns would the dataset have?

1.28 Don't Text While Studying! For the 2015 Intel Science Fair, two brothers in high school recruited 47 of their classmates to take part in a two-stage study. Participants had to read two different passages and then answer questions on them, and each person's score was recorded for each of the two

[17] Sample data obtained from the Census at Schools random sampler sponsored by the American Statistical Association at *https://ww2.amstat.org/censusatschool*.

[18] Data downloaded from the US Department of Education's College Scorecard at *https://collegescorecard.ed.gov/data/* (November 2019.)

[19] Quenqua, D., "Putting a Crimp in the Hookah," *New York Times*, May 31, 2011, p A1.

[20] Sforza, C., Grandi, G., De Menezes, M., Tartaglia, G.M., and Ferrario, V.F., "Age- and sex-related changes in the normal human external nose," *Forensic Science International*, January 30, 2011; 204(1–3): 205.e1–9.

tests. There were no distractions for one of the passages, but participants received text messages while they read the other passage. Participants scored significantly worse when distracted by incoming texts. Participants were also asked if they thought they were good at multitasking (yes or no) but "even students who were confident of their abilities did just as poorly on the test while texting."[21]

(a) What are the cases?

(b) What are the variables? Is each variable categorical or quantitative?

(c) If we create a dataset of the information with cases as rows and variables as columns, how many rows and how many columns would the dataset have?

1.29 Help for Insomniacs A recent study shows that just one session of cognitive behavioral therapy can help people with insomnia.[22] In the study, forty people who had been diagnosed with insomnia were randomly divided into two groups of 20 each. People in one group received a one-hour cognitive behavioral therapy session while those in the other group received no treatment. Three months later, 14 of those in the therapy group reported sleep improvements while only 3 people in the other group reported improvements.

(a) What are the cases in this study?

(b) What are the relevant variables? Identify each as categorical or quantitative.

(c) If we create a dataset of the information with cases as rows and variables as columns, how many rows and how many columns would the dataset have?

1.30 How Are Age and Income Related? An economist collects data from many people to determine how age and income are related. How the data is collected determines whether the variables are quantitative or categorical. Describe how the information might be recorded if we regard both variables as quantitative. Then describe a different way to record information about these two variables that would make the variables categorical.

1.31 Political Party and Voter Turnout Suppose that we want to investigate the question "Does voter turnout differ by political party?" How might we collect data to answer this question? What would the cases be? What would the variable(s) be?

1.32 Wealth and Happiness Are richer people happier? How might we collect data to answer this question? What would the cases be? What would the variable(s) be?

1.33 Choose Your Own Question Come up with your own question you would like to be able to answer. What is the question? How might you collect data to answer this question? What would the cases be? What would the variable(s) be?

[21] Perkins, S., "Studying? Don't answer that text!" *Science News*, July 23, 2015.

[22] Ellis, J.G., Cushing, T., and Germain, A., "Treating acute insomnia: a randomized controlled trial of a 'single-shot' of cognitive behavioral therapy for insomnia," *SLEEP*, 2015; 38(6): 971–978.

1.2 SAMPLING FROM A POPULATION

While most of this textbook is devoted to analyzing data, the way in which data are *collected* is critical. Data collected well can yield powerful insights and discoveries. Data collected poorly can yield very misleading results. Being able to think critically about the method of data collection is crucial for making or interpreting data-based claims. In the rest of this chapter, we address some of the most important issues that need to be considered when collecting data.

Samples from Populations

The US Census is conducted every 10 years and attempts to gather data about all people living in the US. For example, the census shows that, for people living in the US who are at least 25 years old, 84.6% have at least a high school degree and 27.5% have at least a college bachelor's degree.[23] The cases in the census dataset are all residents of the US, and there are many variables measured on these cases. The US census attempts to gather information from an entire *population*. In **AllCountries**,

[23] *http://factfinder.census.gov.*

introduced as Data 1.2 on page 7, the cases are countries. This is another example of a dataset on an entire population because we have data on every country.

Usually, it is not feasible to gather data for an entire population. If we want to estimate the percent of people who wash their hands after using a public restroom, it is certainly not possible to observe all people all the time. If we want to try out a new drug (with possible side effects) to treat cancer, it is not safe to immediately give it to all patients and sit back to observe what happens. If we want to estimate what percentage of people will react positively to a new advertising campaign, it is not feasible to show the ads to everyone and then track their responses. In most circumstances, we can only work with a *sample* from what might be a very large population.

Samples from Populations

A **population** includes all individuals or objects of interest.

Data are collected from a **sample**, which is a subset of the population.

Example 1.10

To estimate what percent of people in the US wash their hands after using a public restroom, researchers pretended to comb their hair while observing 6000 people in public restrooms throughout the United States. They found that 85% of the people who were observed washed their hands after going to the bathroom.[24] What is the sample in this study? What is a reasonable population to which we might generalize?

Solution

The sample is the 6000 people who were observed. A reasonable population to generalize to would be all people in the US. There are other reasonable answers to give for the population, such as all people in the US who use public restrooms or all people in the US who use public restrooms in the cities in which the study was conducted. Also, people might behave differently when alone than when there is someone else in the restroom with them, so we might want to restrict the population to people in a restroom with someone else.

We denote the size of the sample with the letter n. In Example 1.10, $n = 6000$ because there are 6000 people in the sample. Usually, the sample size, n, is much smaller than the size of the entire population.

Since we rarely have data on the entire population, a key question is how to use the information in a sample to make reliable statements about the population. This is called *statistical inference*.

Statistical Inference

Statistical inference is the process of using data from a sample to gain information about the population.

Figure 1.1 diagrams the process of selecting a sample from a population, and then using that sample to make inferences about the population. Much of the data analysis discussed in this text focuses on the latter step, statistical inference. However, the first step, selecting a sample from the population, is critical because the process used to collect the sample determines whether valid inference is even possible.

[24] Zezima, K., “For many, ‘Washroom’ seems to be just a name,” *New York Times*, September 14, 2010, p. A14.

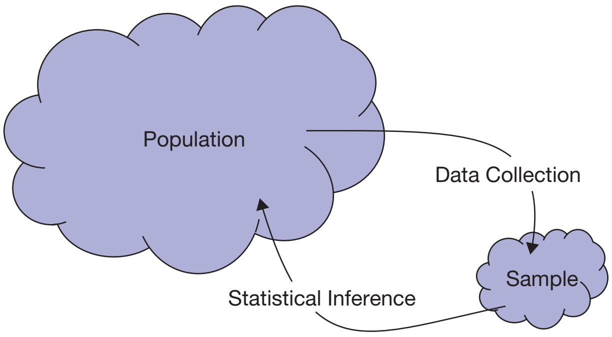


Figure 1.1 ***From population to sample and from sample to population***

Sampling Bias

Example 1.11

Dewey Defeats Truman

The day after the 1948 presidential election, the *Chicago Tribune* ran the headline "Dewey Defeats Truman." However, Harry S Truman defeated Thomas E. Dewey to become the 33rd president of the United States. The newspaper went to press before all the results had come in, and the headline was based partly on the results of a large telephone poll which showed Dewey sweeping Truman.

(a) What is the sample and what is the population?

(b) What did the pollsters want to infer about the population based on the sample?

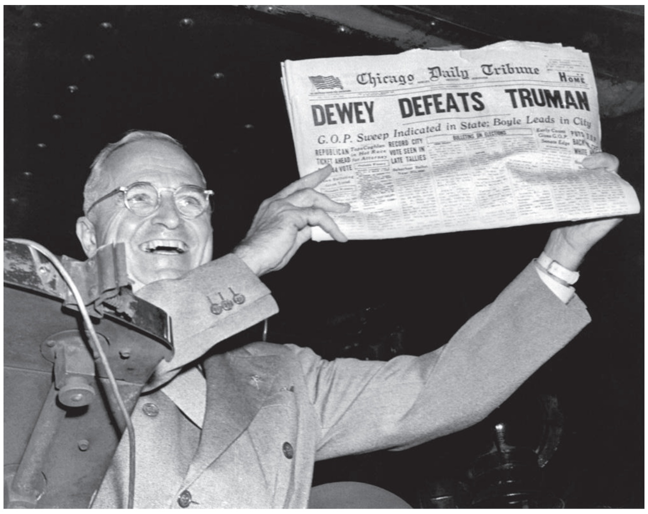
(c) Why do you think the telephone poll yielded such inaccurate results?

Solution

(a) The sample is all the people who participated in the telephone poll. The population is all voting Americans.

(b) The pollsters wanted to estimate the percentage of all voting Americans who would vote for each candidate.

(c) One reason the telephone poll may have yielded inaccurate results is that people with telephones in 1948 were not representative of all American voters. People with telephones tended to be wealthier and prefer Dewey while people without phones tended to prefer Truman.



Underwood Archives/Getty Images

A triumphant Harry S Truman holds the Chicago Tribune published with the incorrect headline "Dewey Defeats Truman"

The previous example illustrates *sampling bias*, because the method of selecting the sample biased the results by selecting only people with telephones.

Sampling Bias

Sampling bias occurs when the method of selecting a sample causes the sample to differ from the population in some relevant way. If sampling bias exists, then we cannot trust generalizations from the sample to the population.

Example 1.12

After a flight, one of the authors received an email from the airline asking her to fill out a survey regarding her satisfaction with the travel experience. The airline analyzes the data from all responses to such emails.

(a) What is the sample and in what population is the airline interested?

(b) Do you expect these survey results to accurately portray customer satisfaction?

Solution

(a) The sample is all people who choose to fill out the survey and the population is all people who fly this airline.

(b) The survey results will probably not accurately portray customer satisfaction. Many people won't bother to fill out the survey if the flight was uneventful, while people with a particularly bad or good experience are more likely to fill out the survey.

A sample comprised of volunteers (like the airline survey) often creates sampling bias in opinion surveys, because the people who choose to participate (the sample) often have more extreme opinions than the population.

To avoid sampling bias, we try to obtain a sample that is *representative* of the population. A representative sample resembles the population, only in smaller numbers. The telephone survey in 1948 reached only people wealthy enough to own a telephone, causing the sample to be wealthier than the population, so it was not a representative sample. The more representative a sample is, the more valuable the sample is for making inferences about the population.

Example 1.13

An online poll conducted on *biblegateway.com* asked, "How often do you talk about the Bible in your normal course of conversation?" Over 5000 people answered the question, and 78% of respondents chose the most frequent option: Multiple times a week. Can we infer that 78% of people talk about the bible multiple times a week? Why or why not?

Solution

No. People who visit the website for *Bible Gateway* and choose to take the poll are probably more likely than the general public to talk about the bible. This sample is not representative of the population of all people, so the results cannot be generalized to all people.

Simple Random Sample

Since a representative sample is essential for drawing valid inference to the population, you are probably wondering how to select such a sample! The key is *random sampling*. We can imagine putting the names of all the cases in the population into a hat and drawing out names to be in our sample. Random sampling avoids sampling bias.

Simple Random Sample

When choosing a **simple random sample** of n units, all groups of size n in the population have the same chance of becoming the sample. As a result, in a simple random sample, each unit of the population has an equal chance of being selected, regardless of the other units chosen for the sample.

Taking a simple random sample avoids sampling bias.

Part of the power of statistics lies in this amazing fact: A simple random sample tends to produce a good representative sample of the population. At the time of writing this book, the population of the United States is more than 300 million people. Although the census collects some data on the entire population, for many questions of interest we are forced to rely on a small sample of individuals. Amazingly, if a simple random sample is selected, even a small sample can yield valid inferences for all 300 million Americans!

Example 1.14

Election Polling

Right before the 2012 presidential election, Google Consumer Surveys[25] randomly sampled and collected data on $n = 3252$ Americans. The sample showed Barack Obama ahead of Mitt Romney in the national popular vote by 2.3 percentage points. Can we generalize these results to the entire population of 126 million voters in order to estimate the popular vote in the election?

Solution

Yes! Because the poll included a random sample of voters, the results from the sample should generalize to the population. In the actual election, Obama won the national popular vote by 2.6 percentage points. Of course, the sample data do not perfectly match the population data (in Chapter 3 we will learn how closely we expect sample results to match population results), but the fact that we can get such an accurate guess from sampling only a very small fraction of the population is quite astonishing!

Analogy to Soup

Patti McConville/Alamy Stock Photo

Sampling the soup

[25] Data from *http://www.fivethirtyeight.com*, "Which Polls Fared Best (and Worst) in the 2012 Presidential Race," November 10, 2012, and Google Consumer Surveys, November 5, 2012, *http://www.google.com/insights/consumersurveys/view?survey=6iwb56wuu5vc6&question=2&filter=&rw=1.*

Do we need to eat an entire large pot of soup to know what the soup tastes like? No! As long as the soup is well mixed, a few spoonfuls will give us a pretty good idea of the taste of the soup. This is the idea behind sampling: Just sampling a small part of the soup (or population) can still give us a good sense of the whole. However, if the soup is not mixed, so that the broth is at the top, the meat at the bottom, and so on, then a few spoonfuls will *not* give us a good sense of the taste of the soup. This is analogous to taking a biased sample. Mixing the soup randomizes the sample, so that the small part we taste is representative of the entire large pot.

How Do We Select a Random Sample?

You may think that you are capable of "randomly" selecting samples on your own, but you are wrong! Deborah Nolan, a statistics professor, has half of her students flip a coin and write the resulting sequence of heads and tails on the board (flipping a coin is truly random), and the other half of her students generate their own sequence of heads and tails without actually flipping the coin, trying to fake randomness. She can always tell the difference.[26] How can she tell? Because students (and their professors and all people!) are very bad at actual randomness.

Similarly, you may think you can select representative samples better than randomness can, but again, you are most likely wrong! Just as humans are surprisingly bad at faking randomness, humans are surprisingly bad at selecting representative samples. We tend to oversample some types of units and undersample less obvious types, even when we are explicitly trying hard not to do so. Luckily, randomness is surprisingly good at selecting a representative sample.

If we can't do randomness ourselves, how *do* we select a random sample? As mentioned, one way is to draw names out of a hat. A more common way is to use technology. Some forms of technology can automatically draw a sample randomly from a list of the entire population, mimicking the process of drawing names from a hat. Other forms produce random numbers, in which case we give a number to each unit in the population, and our sample becomes the units corresponding to the selected numbers.

Example 1.15

The dataset **AllCountries** contains data on 217 countries or economies. Select a random sample of 5 of these.

Solution

There are many ways to do this, depending on the technology or method available. One way is to number the countries from 1 to 217 and then use a random number generator to select five of the numbers between 1 and 217. Suppose we do this and get the numbers

47 88 155 164 50

As we see in the dataset, the corresponding countries are Costa Rica (47), Hungary (88), Peru (155), Samoa (164), and Cuba (50). These five countries are a random sample from the population of all countries. Of course, the numbers are randomly generated, so each sample generated this way is likely to be different. We talk more about the variability of random samples in Chapter 3.

[26] Gelman, A. and Nolan, D., *Teaching Statistics: A Bag of Tricks*, Oxford University Press, New York, 2002.

⚠

Random Sampling Caution

In statistics, random is NOT the same as haphazard! We cannot obtain a random sample by haphazardly picking a sample on our own. We must use a formal random sampling method such as technology or drawing names out of a hat.

Realities of Random Sampling

While a random sample is ideal, often it may not be achievable. A list of the entire population may not exist, it may be impossible to contact some members of the population, or it may be too expensive or time consuming to do so. Often we must make do with whatever sample is convenient. The study can still be worth doing, but we have to be very careful when drawing inferences to the population and should at least try to avoid obvious sampling bias as much as possible.

Example 1.16

If the *Chicago Tribune* had wanted to more accurately predict the outcome of the 1948 presidential election, what should they have done? Why do you think they didn't do this?

Solution

To more accurately predict the 1948 presidential election, they should have selected a random sample from the list of all registered voters and then asked the selected people who they would vote for. There are several possible reasons they did not select a random sample. They may not have had a list of all registered voters available from which to sample randomly. Also, collecting data from a random sample of voters might have required traveling to homes all over the country, which would have been time consuming and expensive. Sampling only people with telephones was cheaper and more convenient.

When it is difficult to take a random sample from the population of interest, we may have to redefine the population to which we generalize.

Example 1.17

What Proportion of People Are Vegetarian?

To determine what proportion of people are vegetarian, we would need to take a random sample of all people, which would be extremely difficult or impossible. How might we redefine our population and question so that it is possible to obtain an accurate estimate?

Solution

One option is to narrow our population to those living in Boston and ask, "What proportion of Bostonians are vegetarian?" It would be possible to take a random sample of telephone numbers from a Boston phone book and call and ask whether they eat meat. In this case our population would only include people with land line phone numbers listed in the Boston phone book so would not include people who rely only on cell phones or who have no phone at all.

For simplicity we only describe a simple random sample in detail, but there are other types of random samples. If we want to know the average weight of a population and want to ensure that the proportion of males and females in our sample matches that of the population, we may take two simple random samples, one within males and one within females. For a study on high school students, it is hard to take a simple random sample. We might first take a simple random sample of schools and then, within each of the sampled schools, take a simple random sample of students. These random sampling schemes are more complicated than the simple random sample but can still yield valid inferences.

Other Sources of Bias

Sampling bias is not the only form of bias that can occur when collecting data. Particularly when collecting data on humans, even if we have a good random sample, there are other issues that might bias the results.

Bias

Bias exists when the method of collecting data causes the sample data to inaccurately reflect the population.

Bias can occur when people we have selected to be in our sample choose not to participate. If the people who choose to respond would answer differently than the people who choose not to respond, results will be biased.

Example 1.18

In 1997 in Somerset (a county in England), a study was conducted on lifestyle choices associated with health.[27] A random sample of 6009 residents of Somerset were mailed a questionnaire that they were asked to fill out and return, and 57.6% of the people in the sample returned the questionnaire. Do you think health-related behavior such as exercise and smoking are accurately portrayed by the data collected?

Solution

Probably not. People who returned the questionnaire may have been more proud of their responses, or may have been more organized and motivated in general, so more likely to lead a healthy lifestyle.

The researchers followed up with phone interviews for a random sample of those who had not responded. As suspected, the non-responders were quite different regarding health behaviors. For example, only 35.9% of initial responders reported getting no moderate or vigorous physical activity, while this percentage was almost doubled, 69.6%, for non-responders. Using only the data from the initial responders is very misleading.

The way questions are worded can also bias the results. In 1941 Daniel Rugg[28] asked people the same question in two different ways. When asked “Do you think that the United States should allow public speeches against democracy?” 21% said the speeches should be allowed. However, when asked “Do you think that the United States should forbid public speeches against democracy?” 39% said the speeches should not be forbidden. Merely changing the wording of the question nearly doubled the percentage of people in favor of allowing (not forbidding) public speeches against democracy.

[27] Hill, A., Roberts, J., Ewings, P., and Gunnell, D., “Non-response bias in a lifestyle survey,” *Journal of Public Health Medicine*, June 1997; 19(2): 203–207.
[28] Rugg, D., “Experiments in wording questions,” *Public Opinion Quarterly*, 1941; 5: 91–92.

iStock.com/Carmen Martínez Banús

Would you have children?

DATA 1.4

"If You Had It to Do Over Again, Would You Have Children?"

In 1976, a young couple wrote to the popular columnist Ann Landers, asking for advice on whether or not to have children.[29] Ann ran the letter from the young couple (which included various reasons not to have kids, but no positive aspects of parenting) and asked her readers to respond to the question "If you had it to do over again, would you have children?" Her request for data yielded over 10,000 responses, and to her surprise, only 30% of readers answered "Yes." She later published these results in *Good Housekeeping*, writing that she was "stunned, disturbed, and just plain flummoxed" by the survey results. She again asked readers to answer the exact same question, and this time 95% of responders said "Yes." ■

Example 1.19

In Data 1.4, why do you think the two percentages, 30% and 95%, are so drastically different?

Solution

The initial request for data was in a column with a letter stating many reasons not to have kids, which may have brought these issues to the minds of the responders. The second request was in an article mentioning Ann Landers' dismay at parents answering no, which may have influenced responses. The context in which the question is asked can bias answers one way or another.

Sampling bias is also present, since readers of her column in the newspaper and readers of *Good Housekeeping* and readers who choose to respond to each request for data are probably not representative of the population and probably differ from each other. For the first request, people with more negative experiences with children may have been encouraged to respond, while the opposite may have been true in the second case. You may be able to think of additional reasons for the discrepancy in the sample results.

Example 1.20

Suppose you are considering having children and would really like to know whether more parents are happy about having kids or regret their decision. Which percentages in Data 1.4 can you trust? How would you go about collecting data you can trust?

[29] *http://www.stats.uwo.ca/faculty/bellhouse/stat353annlanders.pdf.*

Solution

Since both of these samples only include people who decided to write in (volunteer samples) instead of taking a random sample, both almost definitely contain sampling bias, so neither should be trusted. To collect data you can trust, you should take a random sample of all parents (or perhaps take a random sample of all parents of your nationality).

Newsday took a random sample of all parents in the US, asking the same question as in Data 1.4. In this random sample, 91% said "Yes," they would have children again if given the choice. This doesn't mean that exactly 91% of parents are happy they had kids, but because it was a random sample, it does mean that the true percentage is close to 91%. In Chapter 3 we'll learn how to assess exactly how close we expect it to be. (Notice that the initial sample result of 30% is extremely misleading!)

Bias may also be introduced if people do not answer truthfully. If the sample data cannot be trusted, neither can generalizations from the sample to the population.

Example 1.21

Illicit Drug Use

The National Survey on Drug Use and Health[30] selected a random sample of US college students and asked them about illicit drug use, among other things. In the sample, 22.7% of the students reported using illicit drugs in the past year. Do you think this is an accurate portrayal of the percentage of all college students using illicit drugs?

Solution

This may be an underestimate. Even if the survey is anonymous, students may be reluctant to report illicit drug use on an official survey and thus may not answer truthfully.

Bias in data collection can result in many other ways not discussed here. The most important message is to always think critically about the way data are collected and to recognize that not all methods of data collection lead to valid inferences. Recognizing sources of bias is often simply common sense, and you will instantly become a more statistically literate individual if, each time you are presented with a statistic, you just stop, inquire, and think about how the data were collected.

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- Distinguish between a sample and a population
- Recognize when it is appropriate to use sample data to infer information about the population
- Critically examine the way a sample is selected, identifying possible sources of sampling bias
- Recognize that random sampling is a powerful way to avoid sampling bias
- Identify other potential sources of bias that may arise in studies on humans

[30] Substance Abuse and Mental Health Services Administration, Results from the 2009 National Survey on Drug Use and Health: Volume I. Summary of National Findings (Office of Applied Studies, NSDUH Series H-38A, HHS Publication No. SMA 10-4856Findings), Rockville, MD, 2010, *https://nsduhweb.rti.org/*.

Exercises for Section 1.2

SKILL BUILDER 1

In Exercises 1.34 to 1.37, state whether the data are best described as a population or a sample.

1.34 To estimate size of trout in a lake, an angler records the weight of 12 trout he catches over a weekend.

1.35 A subscription-based music website tracks its total number of active users.

1.36 The US Department of Transportation announces that of the 250 million registered passenger vehicles in the US, 2.1% are electro-gas hybrids.

1.37 A questionnaire to understand athletic participation on a college campus is emailed to 50 college students, and all of them respond.

SKILL BUILDER 2

In Exercises 1.38 to 1.41, describe the sample and describe a reasonable population.

1.38 A sociologist conducting a survey at a mall interviews 120 people about their cell phone use.

1.39 Five hundred Canadian adults are asked if they are proficient on a musical instrument.

1.40 A cell phone carrier sends a satisfaction survey to 100 randomly selected customers.

1.41 The Nielsen Corporation attaches databoxes to televisions in 1000 households throughout the US to monitor what shows are being watched and produce the Nielsen Ratings for television.

SKILL BUILDER 3

In Exercises 1.42 to 1.45, a biased sampling situation is described. In each case, give:

(a) The sample

(b) The population of interest

(c) A population we can generalize to given the sample

1.42 To estimate the proportion of Americans who support changing the drinking age from 21 to 18, a random sample of 100 college students are asked the question "Would you support a measure to lower the drinking age from 21 to 18?"

1.43 To estimate the average number of tweets from all twitter accounts in 2019, one of the authors randomly selected 10 of his followers and counted their tweets.

1.44 To investigate interest across all residents of the US in a new type of ice skate, a random sample of 1500 people in Minnesota are asked about their interest in the product.

1.45 To determine the height distribution of female high school students, the rosters are collected from 20 randomly selected high school girls basketball teams.

SKILL BUILDER 4

In Exercises 1.46 to 1.51, state whether or not the sampling method described produces a random sample from the given population.

1.46 The population is incoming students at a particular university. The name of each incoming student is thrown into a hat, the names are mixed, and 20 names (each corresponding to a different student) are drawn from the hat.

1.47 The population is the approximately 25,000 protein-coding genes in human DNA. Each gene is assigned a number (from 1 to 25,000), and computer software is used to randomly select 100 of these numbers yielding a sample of 100 genes.

1.48 The population is all employees at a company. All employees are emailed a link to a survey.

1.49 The population is adults between the ages of 18 and 22. A sample of 100 students is collected from a local university, and each student at the university had an equal chance of being selected for the sample.

1.50 The population is all trees in a forest. We walk through the forest and pick out trees that appear to be representative of all the trees in the forest.

1.51 The population is all people who visit the website *CNN.com*. All visitors to the website are invited to take part in the daily online poll.

IS IT BIASED?

In Exercises 1.52 to 1.56, indicate whether we should trust the results of the study. Is the method of data collection biased? If it is, explain why.

1.52 Ask a random sample of students at the library on a Friday night "How many hours a week do you study?" to collect data to estimate the average number of hours a week that all college students study.

1.53 Ask a random sample of people in a given school district, "Excellent teachers are essential to the well-being of children in this community, and teachers truly deserve a salary raise this year.

Do you agree?" Use the results to estimate the proportion of all people in the school district who support giving teachers a raise.

1.54 Take 10 apples off the top of a truckload of apples and measure the amount of bruising on those apples to estimate how much bruising there is, on average, in the whole truckload.

1.55 Take a random sample of one type of printer and test each printer to see how many pages of text each will print before the ink runs out. Use the average from the sample to estimate how many pages, on average, all printers of this type will last before the ink runs out.

1.56 Send an email to a random sample of students at a university asking them to reply to the question: "Do you think this university should fund an ultimate frisbee team?" A small number of students reply. Use the replies to estimate the proportion of all students at the university who support this use of funds.

1.57 Do Parents Regret Having Children? In Data 1.4 on page 25, we describe the results of a question asked by a national newspaper columnist: "If you had it to do over again, would you have children?" In addition to those results and a follow-up national survey, the *Kansas City Star* selected a random sample of parents from Kansas City and asked them the same question. In this sample, 94% said "Yes." To what population can this statistic be generalized?

1.58 Wearing a Uniform to Work The website *fox6now.com* held an online poll in June 2015 asking "What do you think about the concept of having an everyday uniform for work, like Steve Jobs did?" Of the people who answered the question, 24% said they loved the idea, 58% said they hated the idea, and 18% said that they already wore a uniform to work.

(a) Are the people who answered the poll likely to be representative of all adult workers? Why or why not?

(b) Is it reasonable to generalize this result and estimate that 24% of all adult workers would like to wear a uniform to work?

1.59 Canadians Stream Music In a random sample of 3500 Canadian consumers, about 71% report that they regularly stream music.[31]

(a) Is the sample likely to be representative of all Canadian consumers? Why or why not?

(b) Is it reasonable to generalize this result and estimate that about 71% of all Canadian consumers regularly stream music?

1.60 Climate Change In June 2018, a poll asked a random sample of 1000 US adults whether global warming will be a serious problem for the United States.[32] The results show that 51% think global warming will be a very serious problem, 27% think it will be a somewhat serious problem, and 21% think it will not be a serious problem.

(a) What is the sample? What is the intended population?

(b) Is it reasonable to generalize this result and estimate that 21% of US adults think that global warming will not be a serious problem for the United States?

1.61 Do You Use a Food Delivery App? A 2019 study conducted by eMarketer[33] asked 800 US smartphone users whether they had used a food delivery app at least once in the last month. The survey also asked which food delivery app, if any, was used. The survey showed that 16.3% of respondents had used a food delivery app in the last month. Of those that had used one, 27.6% used DoorDash, 26.7% used Grubhub, 25.2% used UberEats, while the rest used another.

(a) What is the sample? What is the intended population?

(b) What are the cases? What are the variables? Classify variables as quantitative or categorical.

1.62 Pennsylvania High School Seniors Exercise 1.24 describes a dataset, stored in **PASeniors**, for a sample of students who filled out a survey though the US Census at School project. When downloading the sample[34] we specified Pennsylvania as the state and Grade 12 as the school year, then the website chose a random sample of 457 students from among all students who matched those criteria. We'd like to generalize results from this sample to a larger population. Discuss whether this would be reasonable for each of the groups listed below.

[31] "What Moves Today's Teenage Canadian Music Fan?," *http://www.nielsen.com/ca/en/insights/news/2015/what-moves-todays-teenage-canadian-music-fan.html*, Neilsen, Media and Entertainment, June 2, 2015.

[32] *https://www.rff.org/energy-and-climate/surveying-american-attitudes-toward-climate-change-and-clean-energy/*. Accessed July 2019.

[33] "US Food Delivery App Usage Will Approach 40 Million Users in 2019," *www.eMarketer.com*, July 2, 2019. Sample size is approximated.

[34] Sample data obtained Data from U.S. Census at School (*https://www.amstat.org/censusatschool*) used with the permission of the American Statistical Association.

(a) The 457 students in the original sample

(b) All Pennsylvania high school seniors who participated in the Census at School survey

(c) All Pennsylvania high school seniors

(d) All students in the United States who participated in the Census at School survey

1.63 How Easily are You Influenced? Mentally simulate ten tosses of a coin by writing down a sequence of Heads and Tails that might result from ten flips of a fair coin. (Try this now!) When a random sample of people were asked to do this, over 80% of them wrote down Heads as the first flip. Expanding on this result,[35] when the instructions asked for a sequence of "Tails and Heads," participants were more likely to put Tails as the first flip. Indeed, when they were told that an imaginary coin was purple on one side and orange on the other (with the two colors presented in random order), participants were more likely to start with whichever color was mentioned first.[36] Researchers could influence the results just based on the order in which they listed the options.

(a) Is this an illustration of sampling bias or wording bias or both or neither?

(b) If you are letting a friend choose between Option Q and Option W, and you are really hoping that they pick Option W, in what order should you present the options?

1.64 Does Chocolate Milk Come from Brown Cows? The US National Dairy Council, a dairy advocacy group, conducted a survey that appears to show that 7% of Americans (which is about 16.4 million people) believe that chocolate milk comes from brown cows. This result was picked up and shared widely, including by CNN.com, the Washington Post, the Today show, and NPR (National Public Radio).[37] The goal of the survey was to find some fun facts to share and the advocacy group has not made the full survey results, or the sampling method, publicly available.

(a) Do you think it is likely that there is sampling bias in this study? Without knowing how the sample was determined, can we know if it is appropriate to generalize to all Americans? Can you think of a way in which the sample might have been determined that would create sampling bias?

(b) While we don't know how the sample was determined, interviewers on NPR were able to find out how the question was phrased.[38] Survey respondents were asked to select one option to the following: "Where does chocolate milk come from? (a) Brown cows, (b) Black and white spotted cows, (c) I don't know." Do you think the way the question was worded might have biased the results? Give a different possible way to word the question that might give more accurate results.

1.65 How Many People Wash Their Hands after Using the Washroom? In Example 1.10 on page 18, we introduce a study by researchers from Harris Interactive who were interested in determining what percent of people wash their hands after using the washroom. They collected data by standing in public restrooms and pretending to comb their hair or put on make-up as they observed patrons' behavior.[39] Public restrooms were observed at Turner's Field in Atlanta, Penn Station and Grand Central Station in New York, the Museum of Science and Industry and the Shedd Aquarium in Chicago, and the Ferry Terminal Farmers Market in San Francisco. Of the over 6000 people whose behavior was observed, 85% washed their hands. Women were more likely to wash their hands: 93% of women washed, while only 77% of men did. The Museum of Science and Industry in Chicago had the highest hand-washing rate, while men at Turner's Field in Atlanta had the lowest.

(a) What are the cases? What are the variables? Classify each variable as quantitative or categorical.

(b) In a separate telephone survey of more than 1000 adults, more than 96% said they always wash their hands after using a public restroom. Why do you think there is such a discrepancy in the percent from the telephone survey compared to the percent observed?

1.66 Teaching Ability In a sample survey of professors at the University of Nebraska, 94% of them described themselves as "above average" teachers.[40]

[35] Bar-Hillel M, Peer E, Acquisti A, "Heads or tails? A Reachability Bias in Binary Choice," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6), April 28, 2014.

[36] This is called the primacy effect, in which the first option given is more likely to be selected.

[37] Griffin L and Campbell T, "Take that chocolate milk survey with a grain of salt," *The Conversation*, June 28, 2017.

[38] Meikle G, "This survey is as murky as chocolate milk," *Columbia Journalism Review*, June 21, 2017.

[39] Bakalar, "Study: More people washing hands after using bathroom," *Salem News*, September 14, 2010.

[40] Cross, P., "Not can, but *will* college teaching be improved?," *New Directions for Higher Education*, 1977; 17: 115.

(a) What is the sample? What is the population?

(b) Based on the information provided, can we conclude that the study suffers from sampling bias?

(c) Is 94% a good estimate for the percentage of above-average teachers at the University of Nebraska? If not, why not?

1.67 Effects of Alcohol and Marijuana In 1986 the Federal Office of Road Safety in Australia conducted an experiment to assess the effects of alcohol and marijuana on mood and performance.[41] Participants were volunteers who responded to advertisements for the study on two rock radio stations in Sydney. Each volunteer was given a randomly determined combination of the two drugs, then tested and observed. Is the sample likely representative of all Australians? Why or why not?

1.68 What Percent of Young Adults Move Back in with Their Parents? The Pew Research Center polled a random sample of $n = 808$ US residents between the ages of 18 and 34. Of those in the sample, 24% had moved back in with their parents for economic reasons after living on their own.[42] Do you think that this sample of 808 people is a representative sample of all US residents between the ages of 18 and 34? Why or why not?

1.69 Do Cat Videos Improve Mood? As part of an "internet cat videos/photos" study, Dr. Jessica Gall Myrick posted an on-line survey to Facebook and Twitter asking a series of questions regarding how individuals felt before and after the last time they watched a cat video on the Internet.[43] One of the goals of the study was to determine how watching cat videos affects an individual's energy and emotional state. People were asked to share the link, and everyone who clicked the link and completed the survey was included in the sample. More than 6000 individuals completed the survey, and the study found that after watching a cat video people generally reported more energy, fewer negative emotions, and more positive emotions.

(a) Would this be considered a simple random sample from a target population? Why or why not?

(b) Ignoring sampling bias, what other ways could bias have been introduced into this study?

1.70 Diet Cola and Weight Gain in Rats A study[44] fed one group of rats a diet that included yogurt sweetened with sugar, and another group of rats a diet that included yogurt sweetened with a zero-calorie artificial sweetener commonly found in diet cola. The rats that were fed a zero-calorie sweetener gained more weight and more body fat compared to the rats that were fed sugar. After the study was published, many news articles discussed the implication that people who drink diet soda gain more weight. Explain why we cannot conclude that this is necessarily true.

1.71 Armoring Military Planes During the Second World War, the U.S. military collected data on bullet holes found in B-24 bombers that returned from flight missions. The data showed that most bullet holes were found in the wings and tail of the aircraft. Therefore, the military reasoned that more armor should be added to these regions, as they are more likely to be shot. Abraham Wald, a famous statistician of the era, is reported to have argued against this reasoning. In fact, he argued that based on these data more armor should be added to the center of the plane, and NOT the wings and tail. What was Wald's argument?

1.72 Employment Surveys Employment statistics in the US are often based on two nationwide monthly surveys: the Current Population Survey (CPS) and the Current Employment Statistics (CES) survey. The CPS samples approximately 60,000 US households and collects the employment status, job type, and demographic information of each resident in the household. The CES survey samples 140,000 nonfarm businesses and government agencies and collects the number of payroll jobs, pay rates, and related information for each firm.

(a) What is the population in the CPS survey?

(b) What is the population in the CES survey?

(c) For each of the following statistical questions, state whether the results from the CPS or CES survey would be more relevant.

i. Do larger companies tend to have higher salaries?

[41] Chesher, G., Dauncey, H., Crawford, J., and Horn, K., "The Interaction between Alcohol and Marijuana: A Dose Dependent Study on the Effects on Human Moods and Performance Skills," Report No. C40, Federal Office of Road Safety, Federal Department of Transport, Australia, 1986.

[42] Parker, K., "The Boomerang Generation: Feeling OK about Living with Mom and Dad," Pew Research Center, March 15, 2012.

[43] Gall Myrick, J., "Emotion regulation, procrastination, and watching cat videos online: Who watches Internet cats, why, and to what effect?," *Computers in Human Behavior*, June 12, 2015.

[44] Swithers, S.E., Sample, C.H., and Davidson, T.L., "Adverse effects of high-intensity sweeteners on energy intake and weight control in male and obesity-prone female rats." *Behavioral neuroscience*, 2013; 127(2), 262.

ii. What percentage of Americans are self-employed?

iii. Are married men more or less likely to be employed than single men?

1.73 National Health Statistics The Centers for Disease Control and Prevention (CDC) administers a large number of survey programs for monitoring the status of health and health care in the US. One of these programs is the National Health and Nutrition Examination Survey (NHANES), which interviews and examines a random sample of about 5000 people in the US each year. The survey includes questions about health, nutrition, and behavior, while the examination includes physical measurements and lab tests. Another program is the National Hospital Ambulatory Medical Care Survey (NHAMCS), which includes information from hospital records for a random sample of individuals treated in hospital emergency rooms around the country.

(a) To what population can we reasonably generalize findings from the NHANES?

(b) To what population can we reasonably generalize findings from the NHAMCS?

(c) For each of the questions below, indicate which survey, NHANES or NHAMCS, would probably be more appropriate to address the issue.

i. Are overweight people more likely to develop diabetes?

ii. What proportion of emergency room visits in the US involve sports-related injuries?

iii. Is there a difference in the average waiting time to be seen by an emergency room physician between male and female patients?

iv. What proportion of US residents have visited an emergency room within the past year?

1.74 Interviewing the Film Crew on Hollywood Movies There were 1295 movies made in Hollywood between 2012 and 2018. Suppose that, for a documentary about Hollywood film crews, a random sample of 5 of these movies will be selected for in-depth interviews with the crew members. Assuming the movies are numbered 1 to 1295, use a random number generator or table to select a random sample of five movies by number. Indicate which numbers were selected. (If you want to know which movies you selected, check out the dataset **HollywoodMovies**.)

1.75 Sampling Some Starbucks Stores The Starbucks chain has about 24,000 retail stores in 70 countries.[45] Suppose that a member of the Starbucks administration wishes to visit six of these stores, randomly selected, to gather some first-hand data. Suppose the stores are numbered 1 to 24,000. Use a random number generator or table to select the numbers for 6 of the stores to be in the sample.

[45] *https://www.starbucks.com/about-us/company-information/starbucks-company-profile.*

1.3 EXPERIMENTS AND OBSERVATIONAL STUDIES

Association and Causation

Three neighbors in a small town in northern New York State enjoy living in a climate that has four distinct seasons: warm summers, cold winters, and moderate temperatures in the spring and fall. They also share an interest in using data to help make decisions about questions they encounter at home and at work.

- Living in the first house is a professor at the local college. She's been looking at recent heating bills and comparing them to data on average outside temperature. Not surprisingly, when the temperature is lower, her heating bills tend to be much higher. She wonders, "It's going to be an especially cold winter; should I budget for higher heating costs?"
- Her neighbor is the plant manager for a large manufacturing plant. He's also been looking at heating data and has noticed that when the building's heating plant is used, there are more employees missing work due to back pain or colds and flu. He wonders, "Could emissions from the heating system be having adverse health effects on the workers?"

- The third neighbor is the local highway superintendent. He is looking at data on the amount of salt spread on the local roads and the number of auto accidents. (In northern climates, salt is spread on roads to help melt snow and ice and improve traction.) The data clearly show that weeks when lots of salt is used also tend to have more accidents. He wonders, "Should we cut down on the amount of salt we spread on the roads so that we have fewer accidents?"

Each of these situations involves a relationship between two variables. In each scenario, variations in one of the variables tend to occur in some regular way with changes in the other variable: lower temperatures go along with higher heating costs, more employees have health issues when there is more activity at the heating plant, and more salt goes with more accidents. When this occurs, we say there is an *association* between the two variables.

Association

Two variables are **associated** if values of one variable tend to be related to the values of the other variable.

The three neighbors share a desirable habit of using data to help make decisions, but they are not all doing so wisely. While colder outside temperatures probably force the professor's furnace to burn more fuel, do you think that using less salt on icy roads will make them safer? The key point is that an association between two variables, even a very strong one, does not imply that there is a *cause and effect* relationship between the two variables.

Causation

Two variables are **causally associated** if changing the value of one variable influences the value of the other variable.

The distinction between association and causation is subtle, but important. In a causal relationship, manipulating one of the variables tends to cause a change in the other. For example, we put more pressure on the gas pedal and a car goes faster. When an association is not causal, changing one of the variables will not produce a predictable change in the other. Causation often implies a particular direction, so colder outside temperatures might cause a furnace to use more fuel to keep the professor's house warm, but if she increases her heating costs by buying more expensive fuel, we should not expect the outdoor temperatures to fall!

Recall from Section 1.1 that values of an explanatory variable might help predict values of a response variable. These terms help us make the direction of a causal relationship more clear: We say changing the explanatory variable tends to cause the response variable to change. A causal statement (or any association statement) means that the relationship holds as an overall trend—not necessarily in every case.

Example 1.22

For each sentence discussing two variables, state whether the sentence implies no association between the variables, association without implying causation, or association with causation. If there is causation, indicate which variable is the explanatory variable and which is the response variable.

(a) Studies show that taking a practice exam increases your score on an exam.

(b) Families with many cars tend to also own many television sets.

(c) Sales are the same even with different levels of spending on advertising.

(d) Taking a low-dose aspirin a day reduces the risk of heart attacks.

(e) Goldfish who live in large ponds are usually larger than goldfish who live in small ponds.

(f) Putting a goldfish into a larger pond will cause it to grow larger.

Solution

▶ (a) This sentence implies that, in general, taking a practice exam *causes* an increase in the exam grade. This is association with causation. The explanatory variable is whether or not a practice exam was taken and the response variable is the score on the exam.

(b) This sentence implies association, since we are told that one variable (number of TVs) tends to be higher when the other (number of cars) is higher. However, it does not imply causation since we do not expect that buying another television set will somehow cause us to own more cars, or that buying another car will somehow cause us to own more television sets! This is association without causation.

(c) Because sales don't vary in any systematic way as advertising varies, there is no association.

(d) This sentence indicates association with causation. In this case, the sentence makes clear that a daily low-dose aspirin *causes* heart attack risk to go down. The explanatory variable is taking aspirin and the response variable is heart attack risk.

(e) This sentence implies association, but it only states that larger fish tend to be in larger ponds, so it does not imply causation.

(f) This sentence implies association with causation. The explanatory variable is the size of the pond and the response variable is the size of the goldfish.

Contrast the sentences in Example 1.22 parts (e) and (f). Both sentences are correct, but one implies causation (moving to a larger pond makes the fish grow bigger) and one does not (bigger fish just happen to reside in larger ponds). Recognizing the distinction is important, since implying causation incorrectly is one of the most common mistakes in statistics. Try to get in the habit of noticing when a sentence implies causation and when it doesn't.

Many decisions are made based on whether or not an association is causal. For example, in the 1950s, people began to recognize that there was an association between smoking and lung cancer, but there was a debate that lasted for decades over whether smoking *causes* lung cancer. It is now generally agreed that smoking causes lung cancer, and this has led to a substantial decline in smoking rates in the US. The fact that smoking causes lung cancer does not mean that everyone who smokes will get lung cancer, but it does mean that people who smoke are more likely to get it (in fact, 10 to 20 times more likely[46]). Other causal questions, such as whether cell phones cause cancer or whether increasing online advertising would cause sales to increase, remain topics of research and debate. One of the goals of this section is to help you determine when a study can, and cannot, establish causality.

[46] *http://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm#1.*

Confounding Variables

Why are some variables associated even when they have no cause and effect relationship in either direction? As the next example illustrates, the reason is often the effect of other variables.

DATA 1.5

Vehicles and Life Expectancy

The US government collects data from many sources on a yearly basis. For example, Table 1.5 shows the number of vehicles (in millions) registered in the US[47] and the average life expectancy (in years) of babies born[48] in the US every four years from 1970 to 2014. A more complete dataset with values for each of the years from 1970 through 2017 is stored in **LifeExpectancyVehicles**. If we plot the points in Table 1.5, we obtain the graph in Figure 1.2. (This graph is an example of a scatterplot, which we discuss in Chapter 2.) As we see in the table and the graph, these two variables are very strongly associated; the more vehicles that are registered, the longer people are expected to live. ■

Table 1.5 *Vehicle registrations (millions) and life expectancy*

Year	Vehicles	Life Expectancy
1970	108.4	70.8
1974	129.9	72.0
1978	148.4	73.5
1982	159.6	74.5
1986	175.7	74.7
1990	188.8	75.4
1994	198.0	75.7
1998	211.6	76.7
2002	229.6	77.3
2006	244.2	77.7
2010	242.1	78.7
2014	260.4	78.9

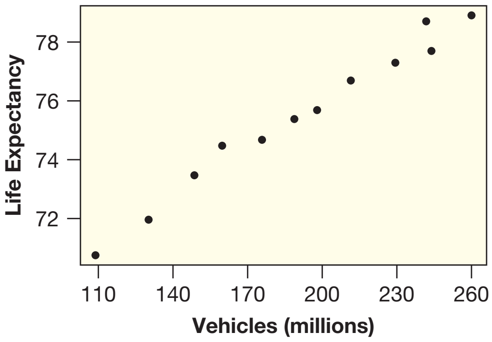


Figure 1.2 *A strong association between vehicles and life expectancy*

[47] Vehicle registrations from US Federal Highway Administration, *https://www.fhwa.dot.gov/policy information/statistics.cfm.*

[48] Centers for Disease Control and Prevention, National Center for Health Statistics, Health Data Interactive, *www.cdc.gov/nchs/hdi.htm*, Accessed October 2019.