STATISTICS

UNLOCKING THE POWER OF DATA

SECOND EDITION

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

WILEY

Statistics

UNLOCKING THE POWER OF DATA

SECOND 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
Miami Dolphins

WILEY

VICE PRESIDENT AND DIRECTOR	Laurie Rosatone
SENIOR ACQUISITIONS EDITOR	Joanna Dingle
DEVELOPMENTAL EDITOR	Adria Giattino
FREELANCE DEVELOPMENTAL EDITOR	Anne Scanlan-Rohrer
EDITORIAL ASSISTANT	Giana Milazzo
SENIOR CONTENT MANAGER	Valerie Zaborski
SENIOR PRODUCTION EDITOR	Laura Abrams
MARKETING MANAGER	John LaVacca
SENIOR PRODUCT DESIGNER	Tom Kulesa
DESIGNER	Thomas Nery
PHOTO EDITOR	Billy Ray
COVER DESIGN	Thomas Nery
COVER PHOTO	© Simone Brandt/Alamy Limited

This book was set in 10/12 TimesTen by SPi Global, and printed and bound by Quad Graphics/Versailles. The cover was printed by Quad Graphics/Versailles.

This book is printed on acid free paper. ∞

Founded in 1807, John Wiley & Sons, Inc. has been a valued source of knowledge and understanding for more than 200 years, helping people around the world meet their needs and fulfill their aspirations. Our company is built on a foundation of principles that include responsibility to the communities we serve and where we live and work. In 2008, we launched a Corporate Citizenship Initiative, a global effort to address the environmental, social, economic, and ethical challenges we face in our business. Among the issues we are addressing are carbon impact, paper specifications and procurement, ethical conduct within our business and among our vendors, and community and charitable support. For more information, please visit our website: www.wiley.com/go/citizenship.

The inside back cover will contain printing identification and country of origin if omitted from this page. In addition, if the ISBN on the back cover differs from the ISBN on this page, the one on the back cover is correct.

ISBN: 978-1-119-30884-3

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

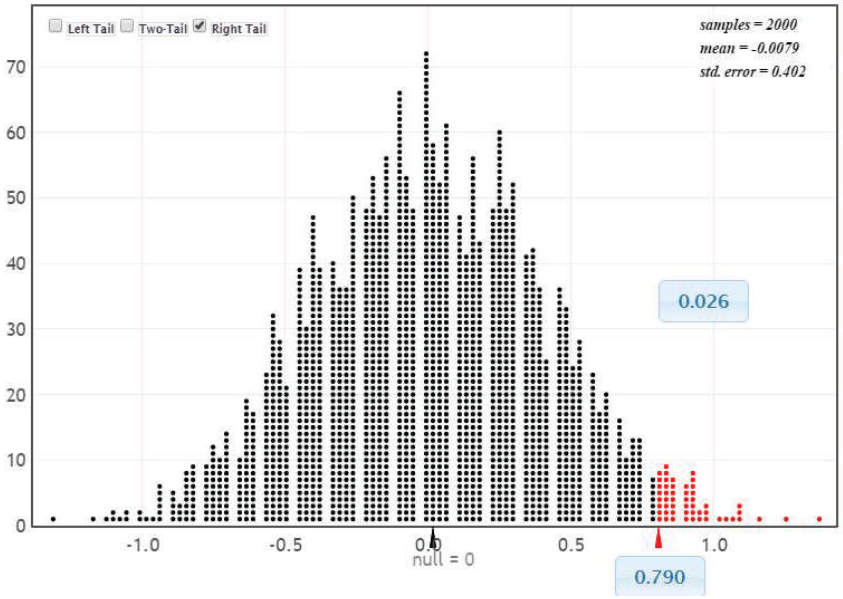
StatKey Randomization Test for a Difference in Means

Leniency and Smiles | Show Data Table | Edit Data | Upload File | Change Column(s)

Randomization method: Reallocate Groups

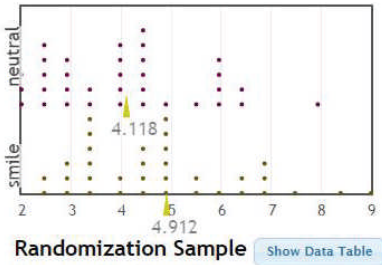
Generate 1 Sample | Generate 10 Samples | Generate 100 Samples | Generate 1000 Samples | Reset Plot

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



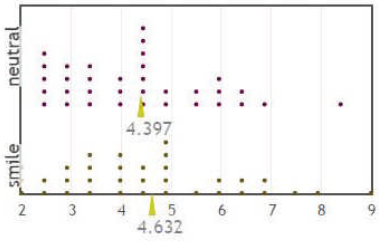
Original Sample

$\bar{x}_1 - \bar{x}_2 = 0.79, n_1 = 34, n_2 = 34$



Randomization Sample | Show Data Table

$\bar{x}_1 - \bar{x}_2 = 0.24, n_1 = 34, n_2 = 34$



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) at

www.wiley.com/college/lock

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.

CONTENTS

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 takes a unique approach of utilizing 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 1800 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. 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*)
- Video solutions for all examples and video tutorials for all learning goals (*www.wiley.com/college/lock*)
- WileyPLUS—an innovative, research-based online environment for effective teaching and learning
- Student solution manual with fully worked solutions to odd-numbered exercises

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 example 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
- Test bank with a wide variety of question types
- 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 single 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.

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 Second Edition

- **Many New Exercises.** The Second Edition includes over 300 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.
- **Many Updated Exercises.** In addition to the many new exercises, this edition also includes over 100 exercises that have been updated with new data.
- **Multiple Variables and Data Visualization.** Chapter 2 has a new section *2.7: Multiple Variables and Data Visualization*, in which we consider a variety of creative and effective ways to visualize data with additional variables and/or data that include geographic or time variables, and illustrate the value of including additional variables.
- **Chapter 4 Reorganized.** Chapter 4 on hypothesis tests has been reorganized to focus more explicitly on one key idea at a time (hypotheses, randomization distributions, p-values, significance) before discussing additional considerations about testing.
- **Chapter 5 Rewritten.** Chapter 5 has been rewritten to improve the transition from Unit B (using simulation methods) to Unit C (using normal and t-based inference).
- **Chapter 6 Sections Re-labeled.** The sections of Chapter 6 have been re-numbered, and some of the sections have been made more concise, to further emphasize the fact that the sections are short and designed to be combined and covered in almost any order.
- **Probability Chapter Renamed.** The Probability chapter has been renamed Chapter P to further emphasize the fact that the chapter is independent of the rest of the material and can be omitted or covered at any point in the course. In addition, a new section has been added to the chapter, on density curves and the normal distribution.
- **StatKey Enhanced.** The online interactive dynamic software *StatKey* has been enhanced, including adding the option to upload whole datasets with multiple variables.
- **And Much More!** We have also made many 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** To aid student learning, we have created video solutions for all examples and short video tutorials for all learning goals. These are available through WileyPLUS or the Student Companion Site. 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, Anne Scanlan-Rohrer, Tom Kulesa, John LaVacca, Laura Abrams, Adria Giattino, Giana Milazzo, Tom Nery, Billy Ray, Valerie Zaborski, 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.

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 a 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, and Ellen Langer. 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.

We appreciate 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. 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.

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 particularly thank Adrian Recinos, who read early drafts of Chapters 3 and 4 to help check that they would be accessible to students with no previous background in statistics.

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 Edition of this book was 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, and Daisy Lock Morgan, and Jocelyn 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
Patti Frazer Lock
Kari Lock Morgan
Eric F. Lock
Dennis F. Lock

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*

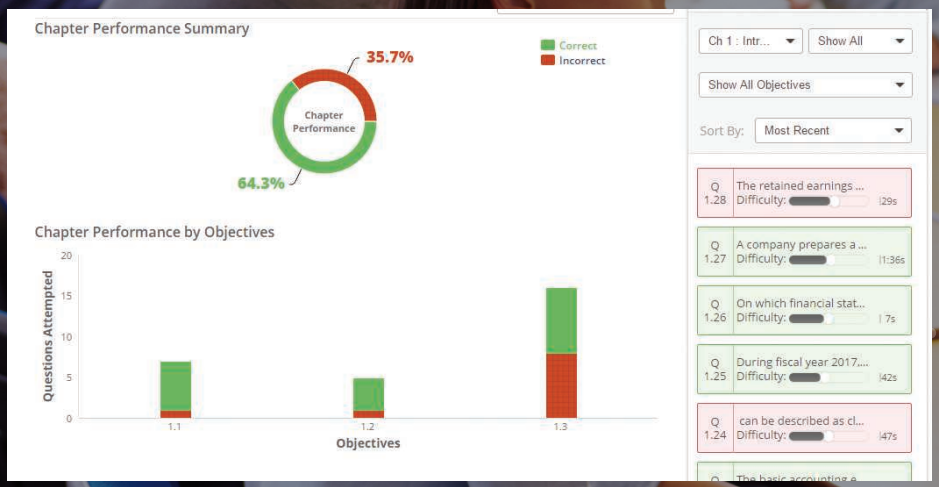
Reviewers and Class Testers for the First Edition

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

WileyPLUS with ORION

A personalized, adaptive learning experience.

WileyPLUS with ORION delivers easy-to-use analytics that help educators and students see strengths and weaknesses to give learners the best chance of succeeding in the course.



Identify which students are struggling early in the semester.

Educators assess the real-time engagement and performance of each student to inform teaching decisions. Students always know what they need to work on.

Help students organize their learning and get the practice they need.

With ORION's adaptive practice, students quickly understand what they know and don't know. They can then decide to study or practice based on their proficiency.

Measure outcomes to promote continuous improvement.

With visual reports, it's easy for both students and educators to gauge problem areas and act on what's most important.

UNIT A

Data

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

New York Times headline, August 5, 2009

UNIT OUTLINE

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.

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*

Top left: ©Pete Saloutos/iStockphoto, Top right: ©Keith Szafranski/iStockphoto, Bottom right: Al Diaz/Miami Herald/MCT via Getty Images

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?
- Does the color red increase the attractiveness of women to men?
- 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?

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 will be generated this 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	Gender	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] *http://www.emc.com/leadership/programs/digital-universe.htm.* Accessed January 2015.

[2] Most datasets used in this text, and descriptions, are available electronically. They can be found at *www.wiley.com/college/lock*. 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:

Gender	M for male and F for female
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.

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, "gender" 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.

- *Gender* is categorical since it classifies students into the two 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:

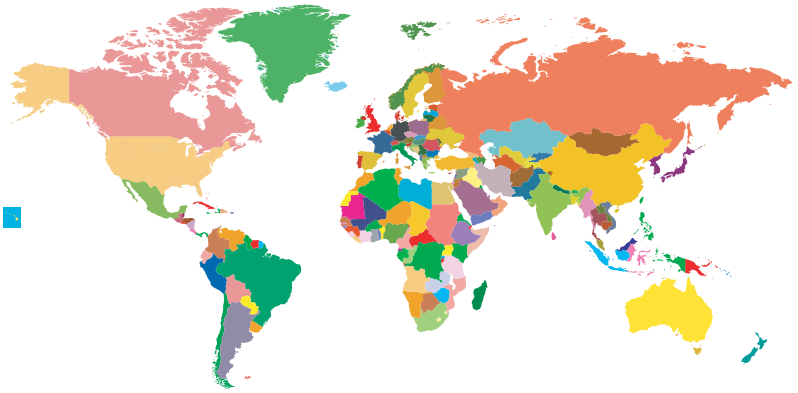
- Who smokes more, males or females?
- Do students who exercise more tend to prefer an Olympic gold medal? Do students who watch lots of television tend to prefer an Academy Award?
- Do males or females watch more television?
- Do students who exercise more tend to have lower pulse rates?
- Do first-borns 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 215 countries listed by the World Bank.[3] 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**. ■



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 Iceland were used to determine that 96.5% of Icelanders have access to the Internet, the highest rate of any country. What are the cases in the data from Iceland? 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 Iceland, the cases are people in Iceland, 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 0.9% in Eritrea to the high of 96.5% in Iceland, and the average is 43.02%. This is a quantitative variable.

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.

[3] *http://data.worldbank.org/indicator/IT.NET.USER.P2.* Data include information on both countries and economies, accessed May 2015.

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.

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[4] 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[5] 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?

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