SECOND EDITION

SOCIAL STATISTICS

Managing Data, Conducting Analyses, Presenting Results

Thomas J. Linneman

ROUTLEDGE

SOCIAL STATISTICS

Many fundamentally important decisions about social life are a function of how well we understand and analyze *data*. This sounds so obvious but it is so misunderstood. Social statisticians struggle with this problem in their teaching constantly. This book and its approach are the ally and a support for all instructors who want to accomplish this hugely important teaching goal.

This innovative text for undergraduate social statistics courses is (as one satisfied instructor put it), a "breath of fresh air." It departs from convention by not covering some techniques and topics that have been in social statistics textbooks for 30 years but that are no longer used by social scientists today. It also *includes* techniques that conventional wisdom has previously thought to be the province of graduate level courses.

Linneman's text is for those instructors looking for a thoroughly "modern" way to teach quantitative thinking, problem-solving, and statistical analysis to their students . . . an undergraduate social statistics course that recognizes the increasing ubiquity of analytical tools in our data-driven age and therefore the practical benefit of learning how to "do statistics," to "present results" effectively (to employers as well as instructors), and to "interpret" intelligently the quantitative arguments made by others.

Thomas J. Linneman is Associate Professor of Sociology at the College of William and Mary in Williamsburg, Virginia. He teaches courses on statistics, social change, sexualities, and the media. At William and Mary, he has been the recipient of the Thomas Jefferson Teaching Award, the highest teaching honor given annually to a younger member of the faculty. The citation for his award noted that Linneman has developed a reputation among his students as a demanding professor but one who genuinely cares about them. His teaching evaluations for his statistics course are regularly a standard deviation above the mean.

Contemporary Sociological Perspectives

Edited by Doug Hartmann, University of Minnesota, Valerie Jenness, University of California, Irvine and Jodi O'Brien, Seattle University

This innovative series is for all readers interested in books that provide frameworks for making sense of the complexities of contemporary social life. Each of the books in this series uses a sociological lens to provide current critical and analytical perspectives on significant social issues, patterns and trends. The series consists of books that integrate the best ideas in sociological thought with an aim toward public education and engagement. These books are designed for use in the classroom as well as for scholars and socially curious general readers.

Published:

Political Justice and Religious Values by Charles F. Andrain
GIS and Spatial Analysis for the Social Sciences by Robert Nash Parker and Emily K. Asencio
Hoop Dreams on Wheels: Disability and the Competitive Wheelchair Athlete by Ronald J. Berger
The Internet and Social Inequalities by James C. Witte and Susan E. Mannon
Media and Middle Class Mom: Images and Realities of Work and Family by Lara Descartes and Conrad Kottak
Watching T.V. Is Not Required: Thinking about Media and Thinking about Thinking by Bernard McGrane and John Gunderson
Violence Against Women: Vulnerable Populations by Douglas Brownridge
State of Sex: Tourism, Sex and Sin in the New American Heartland by Barbara G. Brents, Crystal A. Jackson & Kate Hausbeck
Sociologists Backstage: Answers to 10 Questions About What They Do by Sarah Fenstermaker and Nikki Jones
Gender Circuits by Eve Shapiro
Surviving the Holocaust: A Life Course Perspective by Ronald Berger
Transforming Scholarship: Why Women's and Gender Studies Students Are Changing Themselves and the World by Michelle Berger and Cheryl Radeloff
Stargazing: Celebrity, Fame, and Social Interaction by Kerry Ferris and Scott Harris
The Senses in Self, Society, and Culture by Phillip Vannini, Dennis Waskul and Simon Gottschalk
Who Lives, Who Dies, Who Decides? by Sheldon Ekland-Olson
Surviving Dictatorship by Jacqueline Adams
The Womanist Idea by Layli Maparyan

Social Theory Re-Wired: New Connections to Classical and Contemporary Perspectives by Wesley Longhofer and Daniel Winchester
Religion in Today's World: Global Issues, Sociological Perspectives, by Melissa Wilcox
Life and Death Decisions: The Quest for Morality and Justice in Human Societies, Sheldon Ekland-Olson
Understanding Deviance: Connecting Classical and Contemporary Perspectives, Tammy L. Anderson

Titles of Related Interest:

The Connected City by Zachary Neal
Regression Analysis for the Social Sciences by Rachel A. Gordon
Applied Statistics for the Social and Health Sciences by Rachel A. Gordon
GIS and Spatial Analysis: A Tool for the Social Sciences by Robert Parker and Emily Asencio

This page intentionally left blank

SOCIAL STATISTICS

MANAGING DATA, CONDUCTING ANALYSES, PRESENTING RESULTS

Second Edition

Thomas J. Linneman

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2014
by Routledge
711 Third Avenue, New York, NY 10017

Simultaneously published in the UK
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data

Linneman, Thomas John.
Social statistics : managing data, conducting analyses, presenting results /
by Thomas J. Linneman. — Second Edition.
pages cm. — (Contemporary sociological perspectives)
Includes bibliographical references and index.
1. Social sciences—Statistical methods. 2. Statistics. I. Title.
HA29.L83118 2014
519.5—dc23 2013029016

ISBN: 978-0-415-66146-1 (hbk)
ISBN: 978-0-415-66147-8 (pbk)
ISBN: 978-0-203-07342-1 (ebk)

Typeset in Times New Roman
by Apex CoVantage, LLC

BRIEF CONTENTS

TABLE OF CONTENTS IN DETAIL

This page intentionally left blank

PREFACE TO THE SECOND EDITION

Instructors of introductory social statistics face an unenviable quandary. They want to give their students the skills they need to succeed in the real world of social research, but they realize that if they push their students too far, they risk losing them altogether. Some instructors understandably surrender to this latter concern, opting to teach their students the more basic statistical procedures. Unfortunately, such procedures are seldom used in the real world. If instructors do decide to introduce their students to more contemporary techniques, they encounter course materials that were not developed with the introductory student in mind. This was the position in which I found myself a number of years ago, and I ultimately reached a decision to do something to remedy this dilemma. *Social Statistics: Managing Data, Conducting Analyses, Presenting Results* is my solution. It is the first statistics text that ventures to cover both classic and contemporary techniques in an approachable way that will engage the typical introductory student and make her eager rather than anxious to study this wide array of techniques.

If you compare the table of contents with those of other introductory statistics texts, you will see some similarities and some major differences. The first half of the book contains, on the surface, many of the similarities. The early chapters include many of the topics that one might find in other books: tables and graphs, measures of central tendency and variation, probability distributions, chi-square tests, confidence intervals, *t*-tests, ANOVA, and bivariate regression. I cover these topics innovatively and efficiently in order to prepare the students for the rest of the book. In the second half of the book, students gain significant exposure to a variety of multiple regression techniques that they will find in the real worlds of social research: reference groups, nested modeling, standardized effects, interaction effects, logistic regression, path analysis,

and nonlinearity. In stark contrast to many books with such coverage, I handle these topics at a level that introductory statistics students will find approachable and engaging. For most beginning statistics students in the social sciences, this is the one and only statistics course they will take. If they use *Social Statistics: Managing Data, Conducting Analyses, Presenting Results*, they will leave the course with a strong and varied set of skills that will serve them well as they try to navigate the social science literature or acquire a job.

Although some of these regression techniques may appear in other introductory books, they often do so only as afterthoughts, covered in the most cursory of ways in the final chapter of the book. Unfortunately, this is exactly the point at which students need more explanation, not less. I cover these techniques with a significant—though not overwhelming—level of depth. I explain each technique using unique graphics, visual analogies, and real-world examples. The clear emphasis is on interpretation: given a regression model, or having created one of his or her own, what steps should a student take to make sense of what the model is telling him or her? Combined with their instructor's assistance, this book gets students to the point where they can translate a wide variety of statistical results, whether they are reading social science literature or making a presentation at their job. It guides students through the entire statistical research process: from working with data to get it ready for analysis, conducting the analyses (both by hand and with SPSS), and moving from raw SPSS output to professional presentation. Each chapter ends with graphical, step-by-step SPSS demonstrations, followed by short "from output to presentation" sections that teach students how to present results in clear and compelling ways.

Some instructors may be rightfully dubious about the possibility of introducing their students to some of these techniques. Yet I maintain that, with the help of this book, this is completely possible. I use several strategies to accomplish this. Each chapter includes several simple examples that convey the key aspects of the technique at hand. Most of these examples use data from the General Social Survey (GSS), primarily from 2012, but occasionally from other years. Many chapters contain "interchapter connections" that show how techniques are related to one another, and illustrate that some of the more advanced techniques can be considered extensions of more basic techniques. These connections also help the student through the challenging task of choosing the appropriate technique given a research situation. Each chapter ends with an example or two from the social science literature, showing how social researchers used the chapter's technique in an interesting way. I guide the students through these examples, showing them how to decipher tables that, at first, seem daunting. I make further use of the literature in a unique appendix that features descriptions of 86 social science journal articles from a variety of academic fields. I have vetted these articles, including only those that have statistical results that won't overwhelm introductory students. For each article, I offer a brief description, talk about the techniques the

authors use to make their points (and what pitfalls to watch out for when reading their results), and end with a few questions for the student about the article's use of statistics.

The book emphasizes visual learning in order to make contemporary techniques more approachable. A series of innovative Excel-based live demonstrations and PowerPoint-based animations make many of the techniques come to life. For example, the Excel-based regression demonstration can, in a brief moment, show students the effect of an outlier on a regression line. A PowerPoint animation walks students through one of the book's path models in order to show the power of indirect effects. Instructors are welcome to integrate these demonstrations and animations into their lectures. There are also innovative videos that demonstrate SPSS procedures. These and other helpful instructor support materials (such as detailed answers to all of the end-of-chapter exercises, and a variety of exam questions) can be found on a companion website at URL: **www.routledge.com/cw/Linneman**

For the end-of-chapter exercises, I use more real-world data from five fascinating datasets: the 2012 American National Election Studies, the 2005 World Values Survey, and three datasets form the Pew Internet & American Life Project (on consumption, health, and cyberbullying). Thus, the end-of-chapter exercises are designed for students of varied interests: sociology, political science, marketing, public health, education, criminal justice, and global studies. Here are some examples that illustrate the range of exercise topics:

- With the exercises from the 2012 American National Election Studies, students explore such questions as "Do voters trust the government more than nonvoters?" and "What propels people to be involved in their communities?"
- With exercises from the 2013 PewShop dataset, students explore such questions as "Do smartphones allow people of all ages to engage in technology-enabled shopping experiences?" and "Do income disparities account for technology consumption differences among racial groups?"
- With exercises from the 2012 PewHealth dataset, students explore such questions as "Do men and women use the Internet to seek health information at the same rate?" and "Do people of all ages use Internet information when discussing their healthcare options with their doctors?"
- With exercises from the 2011 PewKids dataset, students explore such questions as "Do children who have been cyberbullied engage in more empathetic behavior toward the cyberbullied than those who have not?" and "What role does parental age play in the level at which parents monitor their children's technology use?"
- With exercises from the 2005 World Values Survey dataset, students explore such questions as "Is the relationship between health and happiness, on a country-by-country level, linear or nonlinear?" and "What role does societal trust play in citizens' desire for authoritarian leadership?"

At every turn, the book gives students opportunities to understand how researchers use social statistics in the real world, and to conduct and present their own analyses, just as they will be expected to do in their own research in academics or employment.

CHAPTERS OF THE BOOK AND WHAT IS NEW TO THIS EDITION

- Chapter 1 is all about forms of data: what do they look like and how do you work with them? Since many students may have never even seen a dataset, I describe how you construct a basic dataset and how you can get it into the shape you want through recoding, computing, and indexing. For example, there is a step-by-step GSS example about constructing an index of workplace hostility. I talk about several of the most innovative and extensive data collection efforts in the social sciences. I discuss how we live in an age of endless data, which presents us with myriad research opportunities, and offer literature examples of researchers using Internet-based data (Wikipedia and the Internet Movie Database) to conduct interesting research projects.
- Chapter 2 covers table construction with one, two, or three variables. I also cover basic graphing, with an emphasis on how to create a graph that accurately represents the data. Examples in this chapter include the effect of childrearing goals on parents' propensity to engage in spanking, and the effects of gender and age on chivalrous behavior. The chapter ends with a fascinating article about racial classification that appeared in a recent issue of a top social science journal, yet featured a greatly exaggerated graph.
- Chapter 3 covers, using a wide variety of unique graphic-based explanations, the basic descriptive statistics: mean, median, mode, variance, and standard deviation. Given that qualitative diversity is of paramount importance, I also provide extensive coverage of the index of qualitative variation, as well as some coverage of the index of dissimilarity and the Gini Coefficient. Examples in this chapter include variation in Internet use by race, and changing attitudes over time toward government spending on health care, the military, and the environment. A literature example highlights the extensive variation in medical costs for a single surgical procedure.
- Chapter 4 is the first of four chapters in the book that cover inferential techniques. In each of these four chapters, I discuss in depth how each technique is based on a probability distribution, showing how such distributions are actually created and what they really mean. In this chapter, I cover inference with crosstabs—the chi-square test—using a creative discussion of statistical significance. I emphasize, through a unique graphic, the effect that sample size can have on chi-square

results. The chapter's examples include the relationship between age and cynicism and the relationships among age, gender, and gun ownership. Both chi-square literature examples involve the body: one covers gender differences in flatulence habits, whereas the other compares how the French and American media treat the obesity epidemic.

- Chapter 5 is the second inference chapter. By hand, I build a sampling distribution and show graphically what the standard error of the mean really is. With regard to applications, this chapter covers testing a population claim and building confidence intervals. Examples in this chapter include attitudes toward police use of violence and the relationship between job stress and job satisfaction. The literature example regards how a researcher used confidence intervals to study how blacks are portrayed in a random sample of contemporary films.
- Chapter 6 is the third inference chapter, and in it, I cover *t*-tests and ANOVA. I construct by hand a sampling distribution of sample mean differences, and I go into significant depth regarding how the tests' formulas actually work. I introduce "interchapter connections," which show students how various techniques are similar or different, and help them understand how to choose among techniques. The examples for the chapter involve the relationships among political party, age, and attitudes toward suicide, and the connection between attitudes toward gender equality in the household and actual behavior within the household. The *t*-test literature example is on gender overcompensation. The ANOVA literature example studies activism through the life course.
- Chapter 7 covers simple bivariate correlation and regression. The graphical examples fully explain the important concept of explained variation. By examining movie grosses over time, I show how regression can be used in forecasting. Other examples include the effects of income on relaxation time, and intergenerational effects on family size. The literature examples cover attitudes toward relinquishing civil liberties in the age of terror, and the correlations among gender, body size, and physical attractiveness.
- Chapter 8 is the final chapter on inference. By building one last sampling distribution, I graphically illustrate what the standard error of the slope represents and how we use it to gauge a regression slope's statistical significance. I emphasize the relationship between sample size and statistical significance, and teach students to think critically about the distinction between statistical and substantive significance. Examples in this chapter examine how level of sexual activity is affected by educational achievement, income, and hours worked. Both literature examples involve grades: looking first at the effect of studying at the college level, second at the effect of family size at the grade-school level.
- Chapter 9 involves the use of various types of variables as independent variables in a regression equation. After covering how to interpret slopes for dichotomous variables, I show in a step-by-step fashion how to use multiple dichotomies to

create a set of reference-group variables. I also include an interchapter connection linking *t*-tests with dichotomous slopes. The examples investigate demographic effects on STEM (science, technology, engineering, and mathematics) achievement, partnership-status effects on happiness, and the relationship between political party and political knowledge. The literature examples show how researchers used dichotomies and reference groups to study gender differences in housework, and temporal changes in attitudes toward gay rights.

- Chapter 10 covers, with the great care that the topic warrants, the very important concept of controlling. I start with some analogies, illustrating how the concept of controlling is actually imbued in our everyday lives. I walk students through the typical tabular construction of a series of nested regression models. I offer an interchapter connection, using the same data to create both an elaborated crosstab and a nested regression model. I show how to judge improvement from model to model, and why it is important to keep sample size constant from model to model. Examples in this chapter involve explaining racial differences in attitudes toward state assistance and gender and in religion's effects on attitudes toward same-sex parenting. The literature examples examine the grade gap between whites and blacks and the media effects on attitudes toward crime.
- Chapter 11 covers the meaning behind standardized coefficients, or betas. Rather than just handing the students the simple formula for calculating betas, I take them through an in-depth explanation so that they can develop a full understanding of what the betas really are and why they are important. I include an interchapter connection the links betas to *z*-scores. Examples involve religiosity and attitudes toward abortion and the male/female differences in what determines life satisfaction. The literature examples cover the topics of school discipline and of country music's effect on suicide rates.
- Chapter 12 covers one of the most prominent techniques in current social science literature: interaction effects. I first make an interchapter connection that illustrates how interaction has similarities to elaborated crosstabs. Then, I show students how to work through examples to develop a full understanding of the interaction. Examples in this chapter examine the interaction effect between sex and number of children on relaxation time, the interaction effect between race and education on attitudes toward Muslim civil rights, and the interaction effect between race and sex on religiosity. Literature examples involve the interaction of gender and religious participation on black political activity, and the interaction of gender and work hours on level of family guilt.
- Chapter 13 explains the difference between regular regression and logistic regression. Without becoming bogged down in the math going on behind the scenes, I show students how to run numerous examples with a logistic regression model in order to understand the probabilities they are calculating. Because so many logistic results are presented as odds ratios, I explain how to interpret such results.

Dichotomous dependent variables in the examples include home ownership, support for gun control, interracial friendships, giving to charity, and condom usage. Literature examples are on the topics of presidential disapproval and global warming.

- Chapter 14 deals with path analysis. Although more esoteric techniques have emerged, I find that path analysis remains a very useful way for students to visualize indirect effects. I describe how to construct and interpret a path model, and in an interchapter connection, I link path analysis and nested models. To this end, I bring back the same-sex parenting example from an earlier chapter and revise it into a path model. There are also examples concerning drinking behavior, political party identification, and intergenerational socioeconomic status effects. The literature examples involve student activism, and emotion work in the service industry.
- Chapter 15 covers simple non-linear relationships and basic log transformations. I include a detailed and graphical explanation of how these nonlinear slopes work. For the examples, I use age's non-linear effect on income, education's nonlinear effect on income, income's non-linear effect on political party, and income's non-linear effect on financial satisfaction. The literature examples involve gendered occupations, and congressional effectiveness.
- Chapter 16 ends the book with a brief look forward, telling students what they might want to look out for as they enter the world of social research. I offer examples of two common regression-related problems: outliers and multicollinearity. Then I very briefly introduce several common advanced techniques that they might encounter in the social research literature, techniques used for specific types of variables (ordered logistic, multinomial), types of samples (multilevel modeling), and types of situations (structural equation modeling, hazard modeling).

I began this preface with a longstanding problem: many of our introductory statistics students do not gain exposure to the techniques they need to know. At academic conferences, from individual discussions to packed workshops on how to transform the introductory statistics course, I have witnessed concern about this situation. Many instructors want to make this type of change, but they simply haven't known how to accomplish it. *Social Statistics: Managing Datasets, Conducting Analyses, Presenting Results* provides instructors with a proven way to achieve this change in their courses. The book markedly improved my own course: I was able to help my students achieve a greater level of understanding of these techniques than ever before. From their reduced stress levels over the material to the improved quality of their class presentations, I witnessed positive change in a number of important ways. I also have heard from other instructors who used the book that students have responded very positively to it and that it has improved their courses. If given the right tools, instructors can teach

their students these contemporary techniques. I believe such changes in the introductory social statistics course are not only possible but also necessary in our data-filled world. We must give students the foundation they need to succeed in their courses, their research, their jobs, and their lives. It is my sincere belief that this book will help us accomplish these goals.

ACKNOWLEDGMENTS

First, I'd like to thank the many students in my statistics courses, at the University of Washington and at the College of William and Mary, who helped me through the years craft my teaching of statistics. I'd particularly like to thank students in my recent courses. As this book became a reality, and as I began working on the new edition, they were regular audiences for various examples I was trying, and they were refreshingly candid about what worked and what did not. My research assistants were invaluable to this project. For the first edition, Margaret Clendenen was a big help with getting the datasets in order. For the second edition, Sarah Overton took on every task I threw at her and returned high quality results with amazing speed. The article and variable appendices in particular are much better because of her. My colleagues were, as always, willing to help, particularly Salvatore Saporito and Graham Ousey.

Steve Rutter at Routledge has been an amazing editor. For the second edition in particular, he has provided me with endless and savvy input. Although I occasionally dreaded the arrival of our phone meetings, by the end of them I always felt reinvigorated. Editorial assistant Margaret Moore expertly kept track of every detail, and had the patience of a saint. Project Manager Deepti Agarwal was a pleasure to work with during the production process. I'd also like to thank Series Editors Val Jenness and Jodi O'Brien for their continuing support and advice. The numerous reviewers of the previous edition offered advice with impressive attention to detail, and the book is all the better for it: thanks to Nathan Wright, Yingyi Ma, Michael Abel, Amy Stone, Melanie Arthur, Sally Raskoff, Matthew Green, Dawn Baunach, Linda Henderson, David Sikkink, Mark Handcock, and Matt Huffman. Also thanks to the reviewers of this edition:

Michael Henderson	University of Mississippi
Veena Kulkarni	Arkansas State University

Michael Stern	College of Charleston
Claude Rubinson	University of Houston, Downtown
Justin Berg	University of North Dakota
B. Mitchell Peck	University of Oklahoma
David Merolla	Wayne State University
Sachi Ando	Widener University
Joseph Baker	East Tennessee State University
Tetsuya Matsubayashi	University of North Texas
Paul Warwick	Simon Fraser University
Charles Kaylor	Temple University
Matt Huffman	University of California, Irvine

I wrote most of this book at my home in Richmond, Virginia, often surrounded by pups and within earshot of my partner Farhang. The pups—Miss(ed) Sunshine, Mistah Jack, and Stanley—provided important reality checks ("Sure interaction effects are important, but we want to interact with *you*"). Farhang (who brought me a cup of coffee just moments ago!) has provided a trophy-worthy level of support. When the book revision took on a life far larger than either of us had imagined, we buckled down and got through it together.

Chapter 1

LIFE IN A DATA-LADEN AGE: FINDING AND MANAGING DATASETS

This chapter covers . . .

. . . what data look like in their raw form within a dataset
. . . how to work with data to get them ready to analyze
. . . the wide variety of datasets that are readily available for analysis
. . . the newer forms that data take, from Internet databases to media analyses
. . . types of variables used in statistical analysis
. . . a classification of statistical procedures we'll cover in this book
. . . examples of how researchers used Wikipedia and IMDb to conduct studies

INTRODUCTION

Well, here we are. Long pause. Awkward silence. Let's get one thing out in the open right away: "thrilled" might not be a good description of your mood at this very moment. You are not thrilled to be sitting in front of a book on statistics. Other emotions likely are in play: boredom, trepidation, fear. Maybe not all of these, but if you're like many students taking a course in statistics, the probability is high that some of these emotions are involved. Any effort I make here to dispel such emotions likely will elicit another set of reactions: skepticism, disbelief, anger at my patronizing tone. I realize it might take me a while to win you over. But I will do my best. Mark my words: at some point, perhaps not right away, but somewhere down the road, you will, perhaps secretly, start to like statistics.

OK, you may not get to that point. But I do hope to convince you that understanding statistics is completely possible if you have the right combination of guides (your instructor and me). It is not only possible to understand statistics; it is also absolutely *essential* to being an informed and effective citizen, activist, or employee. We live in an age in which information is overwhelmingly everywhere, and a lot of this information is statistical. Legislators measure the success of social policies based on statistics. A philanthropist considering funding a nonprofit organization may ask for evidence of the organization's prior success, and this evidence is often statistical in nature. Start-up companies have made fortunes by developing better statistical models to help people mine the data created daily by people's Internet searches and by consumer behavior. Therefore, if you can't speak statistics, or read them, you could very well be left out of the loop.

Did I just say, "speak statistics"? Yes, I did. In many ways, for many people, learning statistics is very similar to learning a foreign language. If I started speaking, say, Farsi or Swahili right now, I'd probably lose your interest rather quickly (unless, of course, you're a speaker of these languages). But do I lose you any less slowly when I say. "Adding the squared age term raises the explained variation by 0.04 (with an F-test significant at $p < .01$) and causes the interaction term to lose its statistical significance?" I'd bet not. Right now, to figure out what this sentence meant, you'd need to take it to someone who speaks statistics, and you'd be relying on that person's translation. By the end of this book, you'll be able to figure out on your own what such sentences mean, which means that, among your friends, family, and coworkers, *you* will likely become the statistical translator. And those statistical tables you see in academic journals or policy briefings? You know, those tables that you just skip over because you have no idea what they're saying? I'll be giving you the necessary skills to be able to navigate such tables with ease.

This book differs substantially from other introductory statistics books. I think that's a good thing, but, granted, I'm biased. In addition to using a writing style I hope will not bore or confuse you, I get us through the basic statistics relatively quickly. I do this in order to spend much more time than most books do on the statistical techniques that are used most in the real world. In my opinion, many books spend far too many chapters going over statistical techniques that students likely will never see in practice. Then, before they get to the really good stuff, the book ends. This is akin to a movie that has lots of character and plot development, and then, right at the climax, when the school bus filled with orphans is hanging off the cliff, the screen fades to black and the credits roll. This book, in contrast, not only saves those orphans; it finds them all families and buys each child a puppy. In this book, I cover the basics and then get to the good stuff. Although I've done my best to write as clearly as possible, there inevitably will be points where, the first time you read through them, something just doesn't

make sense. Don't give up there. Sometimes this material takes a few readings before you really understand it. But, if you are persistent, you will get there.

WHAT DATA LOOK LIKE

Yes, *look*. The word *data* is the plural form of the singular word *datum*. It may sound weird now, but get used to it, because it's grammatically correct. Stratum, medium, datum; strata, media, data. The data *are* correct. The data *are* available on the Internet. The data *do* not lie. Actually, sometimes they do lie, but more on that later in the book. In our trip together, we'll be calculating and interpreting statistics using lots and lots of data, so the first things I want to go over with you are the basic forms that data take, the major sources of data today, and some useful ways to work with data to answer the questions you want to answer. Here's a hypothetical short conversation between me and a computer:

TL: Hello, computer, I'm a male.
Computer: 00010110110001001?
TL: I am a male.
Computer: 00010110110001001?
TL: (sigh) 01100001010001!
Computer: 01100001010001? 010011011!!!
TL: 011011000011001.

Computers, as amazing as they are, don't understand words very well. Of course, we're getting there; voice recognition is no longer just a dream. But, even with such programs, behind the scenes the computer still is using numbers. Data in the social sciences, then, are almost always reduced to numbers. However, when researchers collect data, it is often through interviews or surveys. We start with a survey interviewer collecting data from a survey respondent. Next, that respondent's answers are translated into numerical codes that the researchers then input into a dataset. The researchers then use the dataset and a statistical program to calculate their statistics. Reducing people's complex behaviors and attitudes to numbers is not a perfect process. Interesting details sometimes get lost in translation. I'll be the first to defend those who use more qualitative techniques to study the social world. However, because this *is* a book on statistics, we'll be working with the more quantitative, survey-driven data.

Before we look at some real datasets, let's start hypothetically, on a very small scale. We conduct a survey of a whopping six people, asking them the following five questions:

1. Their sex (male or female)
2. Their age (in years)
3. Their race (white, black, or other)
4. The highest level of education they have completed (some high school, high school diploma, some college, college degree, advanced degree)
5. Their support of capital punishment for someone convicted of murder (strongly support, support, oppose, strongly oppose).

Here are the tiny surveys for each person:

Exhibit 1.1: A Tiny Set of Data

Respondent 1: 1. Male 2. 42 3. White 4. High school diploma 5. Strongly supports	**Respondent 2:** 1. Male 2. 75 3. Other 4. College degree 5. Opposes	**Respondent 3:** 1. Female 2. 20 3. White 4. Some high school 5. Supports
Respondent 4: 1. Male 2. 56 3. Black 4. Advanced degree 5. Strongly opposes	**Respondent 5:** 1. Female 2. 33 3. White 4. College degree 5. No answer	**Respondent 6:** 1. Female 2. 63 3. Black 4. High school diploma 5. Strongly opposes

We have data! Now what do we do with them? If we're like most people, we'll enter them into a statistical analysis program. In this book, we'll use SPSS, because it is one of the easiest to use and it is one of the most widely used statistical packages. Most chapters in this book end with a series of SPSS demonstrations that cover the statistical techniques I just went over in the chapter. But, for now, we'll stay hypothetical and get to SPSS at the end of the chapter. First, we need to name our variables. The first four are easy to name: SEX, AGE, RACE, DEGREE. We could name the last one CAPITAL PUNISHMENT, but typically variable names are shorter, so we'll go with CAPPUN. With those decisions made, our dataset looks like this:

Exhibit 1.2: An Empty Dataset

	SEX	AGE	RACE	DEGREE	CAPPUN
1					
2					
3					
4					
5					
6					

Each variable gets its own column, and each respondent gets his or her own row. Now we need to fill in the cells with the data. For the age variable, we can just put in the actual numbers. However, because the computer doesn't like words, we next need to assign numbers, or **codes**, for each category of the other four of our variables. For SEX, with two categories, we'll use

Male: 0, Female: 1

Now, men, don't ascribe too much meaning to this. I don't think you're zeros. Coding is often arbitrary. I just as easily could have coded females as 0 and males as 1. For RACE, with three categories, we'll use

White: 0, Black: 1, Other: 2

For DEGREE, with five categories, we'll use

Some high school: 0, High school diploma: 1, Some college: 2, College degree: 3, Advanced degree: 4

Finally, the capital punishment variable has four categories:

Strongly support: 0, Support: 1, Oppose: 2, Strongly oppose: 3

With the codes in place, we can fill in our cells:

Exhibit 1.3: A Filled-In Dataset

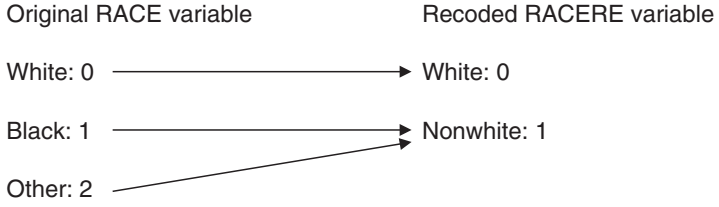
	SEX	AGE	RACE	DEGREE	CAPPUN
1	0	42	0	1	0
2	0	75	2	3	2
3	1	20	0	0	1
4	0	56	1	4	3
5	1	33	0	3	.
6	1	63	1	1	3

Be sure to observe how the codes match up with the respondents' survey answers. Notice that Respondent 5 did not give an answer to Question 5, so she gets a dot for that variable (you'll see this a lot, or people will designate particular numbers for "no answer" or "don't know"). Our little dataset, with five variables and six respondents, has 30 cells with 30 pieces of information. As you likely can imagine, most datasets have many more respondents and variables than this one. For example, over the years, the General Social Survey (GSS) has interviewed 57,061 respondents and has 5,548 variables, giving us 316,574,428 cells. That's a lot of information.

MAKING THE DATA WORK FOR YOU

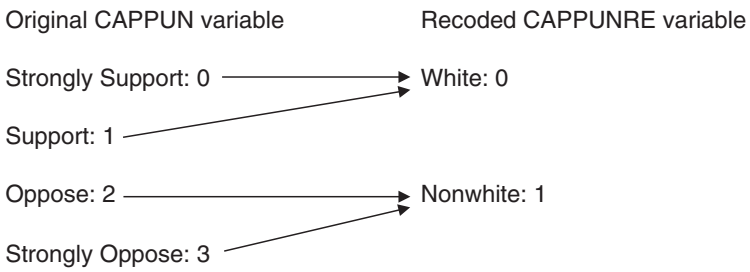
Most of the time, the data aren't exactly in the shape that we need them to be. But this is easy enough to fix. What we do is take the original variable and go through a process called **recoding** in order to create a new variable (always leaving the old variable in its original state). If we wanted to compare whites and nonwhites with regard to whether they either support or oppose capital punishment, we'd want to recode each of these variables. First, we take the original RACE variable and recode it from three categories to two categories, giving us a new variable we'll call RACERE:

Exhibit 1.4: Recoding the RACE Variable



This gives us a new variable with two categories instead of the original three. Next, we take the CAPPUN variable and from it recode a new variable we'll call CAPPUNRE:

Exhibit 1.5: Recoding the CAPPUN Variable



This gives us a new variable with two categories instead of the original four. When we take an original variable and from it create a new variable with fewer categories, we call this process **collapsing**: we are collapsing multiple categories into a smaller number of categories. Here is our dataset with the two new variables and their values:

Exhibit 1.6: Our Tiny Dataset with Two New Variables

	SEX	AGE	RACE	DEGREE	CAPPUN	RACERE	CAPPUNRE
1	0	42	0	1	0	0	0
2	0	75	2	3	2	1	1
3	1	20	0	0	1	0	0
4	0	56	1	4	3	1	1
5	1	33	0	3	.	0	.
6	1	63	1	1	3	1	1

Notice that when we collapse a variable, we do lose valuable detail. For example, with the new CAPPUNRE variable, we no longer know if someone strongly opposes or simply opposes capital punishment. Therefore, as a general rule, we should collapse categories together only when we have good reason to do so. Don't just collapse willy-nilly just because it's fun. A good reason could be substantive: we really want to compare whites to nonwhites. Or our reason could be statistical: to do what we want to do statistically, we need the variable to be collapsed.

Another way to get new variables is to combine variables in a variety of ways. What if we were studying the household division of labor (a fancy way of saying who does the housework), and we had these two measures:

YOURHWK: number of hours of housework the respondent does per week
PRTNRHWK: number of hours of housework the respondent's partner does per week

We could take these measures and calculate other variables:

TOTALHWK = YOURHWK + PRTNRHWK: this would give us the combined hours of housework.
YOURHWKPCT = YOURHWK/TOTALHWK: this would give us the proportion of housework the respondent does.
PRTNRHWKPCT = PRTNRHWK/TOTALHWK: this would give us the proportion of housework the respondent's partner does.

Another way to combine variables is to find related variables and engage in a process called **indexing**. The simplest (and most common) form of index is an additive index, in which we add variables together. Let's say we had our six respondents from earlier respond not to a single question about capital punishment but to three questions with specific scenarios (e.g., Would you support capital punishment for terrorists responsible for 9/11?), where 0 = NO and 1 YES. We could add respondents' responses together into an index called CAPPUNDX. The dataset might look like this:

Exhibit 1.7: A Dataset with an Additive Index

	CAPPUN1	CAPPUN2	CAPPUN3	CAPPUNDX
1	0	0	0	0
2	1	1	1	3
3	1	0	0	1
4	0	1	1	2
5	1	1	.	.
6	0	0	0	0

Ornery Respondent 5 is at it again: she was willing to answer the first two questions, but not the third. So, unfortunately, we are unable to compute an index score for her.

Some people use techniques to take care of this, such as "guessing" what the respondent would have said for the question she skipped, based on how she answered other questions, but I'm not a big fan of this.

Indexes (or indices) are a great idea if you're worried that a single variable will not adequately capture what you're trying to get at. Indexing is a real art, and some people get really picky about how you do it, but we won't go into such detail in this book. Just make sure your combinations of variables make sense, not just to you but also to any "reasonable" person. For example, in our original five-variable dataset, if someone said, "Let's add together the age and degree variables," we'd want to meet such a suggestion with great skepticism. Another question we want to ask ourselves when we create an additive index is, "Do our variables give us enough combined cases?" For example, what if our CAPPUN1,2,3 dataset looked like this:

Exhibit 1.8: An Unfortunate Attempt at an Additive Index

	CAPPUN1	CAPPUN2	CAPPUN3	CAPPUNDX
1	.	0	0	.
2	.	1	1	.
3	.	0	0	.
4	0	1	.	.
5	1	1	.	.
6	0	0	.	.

So very sad. It seems that the survey researchers did not plan things out well. They didn't ask any of the respondents all three CAPPUN questions. So our index ends up with no respondents with valid index scores. If your index involves several variables, sometimes there's just one variable that turns out to be the culprit (later in the book, we'll go over how to figure this out). Removing this variable from the index would likely solve the problem.

OUR DATASETS

In this book, I'm using data from six well-known datasets. For the in-chapter examples, I use the General Social Survey (GSS). For the end-of-chapter exercises, I use