

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

© 2014 Taylor & Francis

The right of Thomas J. Linneman to be identified as author of this work has been asserted by him in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilized in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

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

<i>Preface</i>	<i>xxix</i>
<i>Acknowledgments</i>	<i>xxxvii</i>
Chapter 1: Life in a Data-Laden Age: Finding and Managing Datasets	1
Chapter 2: The Art of Visual Storytelling: Creating Accurate Tables and Graphs	46
Chapter 3: Summarizing Center and Diversity: Basic Descriptive Statistics	92
Chapter 4: Using Sample Crosstabs to Talk about Populations: The Chi-Square Test	141
Chapter 5: Using a Sample Mean or Proportion to Talk about a Population: Confidence Intervals	189
Chapter 6: Using Multiple Sample Means to Talk about Populations: <i>t</i> -Tests and ANOVA	231
Chapter 7: Give Me One Good Reason Why: Bivariate Correlation and Regression	264
Chapter 8: Using Sample Slopes to Talk about Populations: Inference and Regression	303
Chapter 9: It's All Relative: Dichotomies as Independent Variables in Regression	326
Chapter 10: Above and Beyond: The Logic of Controlling and the Power of Nested Regression Models	348

Chapter 11:	Some Slopes Are Bigger than Others: Calculating and Interpreting Beta Coefficients	384
Chapter 12:	Different Slopes for Different Folks: Interaction Effects	402
Chapter 13:	Explaining Dichotomous Outcomes: Logistic Regression	435
Chapter 14:	Visualizing Causal Stories: Path Analysis	467
Chapter 15:	Questioning the Greatness of Straightness: Nonlinear Relationships	493
Chapter 16:	Problems and Prospects: Regression Diagnostics, Advanced Techniques, and Where to Go Now	532
<i>Appendix A:</i>	<i>Variables and Indexes from the Datasets Used in the End-of-Chapter Exercises</i>	<i>A-1</i>
<i>Appendix B:</i>	<i>86 Articles That Use Statistics in Less Than Scary Ways</i>	<i>B-1</i>
<i>Appendix C:</i>	<i>Statistical Tables</i>	<i>C-1</i>
<i>Appendix D:</i>	<i>Answers to Selected End-of-Chapter Exercises</i>	<i>D-1</i>
	<i>Bibliography</i>	<i>R-1</i>
	<i>Glossary/Index</i>	<i>I-1</i>

TABLE OF CONTENTS IN DETAIL

<i>Preface</i>	<i>xxix</i>
<i>Acknowledgments</i>	<i>xxxvii</i>
Chapter 1: Life in a Data-Laden Age: Finding and Managing Datasets	1
Introduction: details the ways in which this book is unique	1
What Data Look Like: introduces the forms that data take in the rawest of ways, and what variables and categories are	3
Making the Data Work for You: covers the very important ways in which we manage data, such as recoding and indexing	6
Our Datasets: describes the datasets that are used throughout the book in examples and exercises: the General Social Survey, the American National Election Studies, the World Values Survey, and surveys from the Pew Internet & American Life Project	9
Other Great Datasets: briefly describes other commonly used datasets, such as the Panel Study of Income Dynamics, and the National Longitudinal Study of Adolescent Health	11
GSS Example: An Index of Work Hostility: illustrates how to recode and combine variables to develop an index of the hostility workers experience at their jobs	13
New Forms of Data: makes the case that there are many new forms of data available for analysis due to the explosion of media availability, old and new	17
Levels of Measurement: carefully delineates the differences among different types of variables and why knowing these differences are important	18

	Major Types of Statistical Procedures: covers the three types of statistics that the book covers: descriptive, inferential, and explanatory	22
	Literature Example: Wikipedia as a Data Source: describes a piece of research that argues that wiki-led organizations play by different rules than most organizations, and uses contested Wikipedia pages to make the argument	23
	Literature Example: IMDb as a Data Source: provides an example of how researchers used the vast database of movies, actors, and directors to study social networks	24
	Conclusion	25
	SPSS Demonstrations: introduces students to SPSS, and covers very important ways to work with data, such as recoding and computing new variables	25
	From Output to Presentation: recommends what to say about a dataset when you're using it in a presentation or paper	40
	Exercises: examine such topics as American attitudes toward wiretapping, concern over China's economic expansion, and social desirability bias in a survey of teenagers	41
Chapter 2:	The Art of Visual Storytelling: Creating Accurate Tables and Graphs	46
	Introduction: makes the case for the importance of tables and graphs, and knowing how to make good ones that don't lie	46
	Tables with One Variable: Frequency Distributions: takes students step by step through the process of building a clear summary of a single variable	47
	GSS Example: Number of Children: walks students through a typical frequency distribution from SPSS, such as the difference between percents and valid percents	48
	Grouped Frequency Distributions: describes when it might be preferable to present a variable's data in grouped form	49
	Tables with Two Variables: Crosstabulations: offers a step-by-step demonstration of how to construct a table with an independent variable and a dependent variable	50
	GSS Example: Spanking and Child-Rearing Goals: uses crosstabs to show how Americans' propensity to spank their children is related to specific goals	53
	GSS Example: Education and Internet Access: models for students every possible mistake you could make, resulting in a horrible crosstab; then reconstructs it correctly	55

Tables with Three Variables: Elaboration in Crosstabs: covers the very important process of elaborating a relationship, showing where a relationship is stronger or weaker, exists or doesn't exist	57
GSS Example: Chivalry, Age, and Gender: uses elaboration to examine who is most likely to give up a seat for another person	58
GSS Example: Racial Differences over Time: offers an example of how a relationship changes over time, using elaboration to show this	60
GSS Example: Gender, Unemployment, and Happiness: investigates whether men or women are more affected by unemployment	61
Graphs with One Variable: examines how to choose among bar, line, and pie graphs	63
Graphs with Two Variables: covers clustered bar graphs and stacked bar graphs	66
Graphs with Three Variables: covers 3-D bar graphs and plotted pie graphs	68
Tufte's Lie Factor: offers a classic way to judge the extent to which a graphic accurately represents the change in data	70
GSS Example: Support for Marijuana Legalization: compares two bar graphs that represent the same data in different ways, leading to different emphases	73
Literature Example: Changing Racial Classification: describes an article in a top research journal that presented a misleading graph	75
Conclusion	77
SPSS Demonstrations: goes through table and graph construction in numerous ways	77
From Output to Presentation: shows how crosstabs don't usually appear in pure crosstab form in professional settings	86
Exercises: explore such issues as perceptions of racial discrimination, the effect of caring for a sick child on one's quality of life, and how many parents say their children don't use social networking sites, when actually they do	87
Chapter 3: Summarizing Center and Diversity: Basic Descriptive Statistics	92
Introduction: discusses in general the field of descriptive statistics	92
Three Ways of Thinking about the Center: using a hypothetical example of housing prices, covers the mean, median, and mode	93
GSS Example: TV Watching: gives an example of how, just based on a few descriptive statistics, we can develop a pretty good guess regarding what the distribution for a variable looks like	96

Procedures for Finding the Median: covers the basics of median finding with even and odd numbers of cases	97
Finding the Centers with a Frequency Distribution: goes through the procedures for locating the mean, median, and mode when the data are in frequency distribution format	98
Measures of Center and Levels of Measurement: explains the connections among descriptive statistics and the levels of measurement covered in Chapter One	100
Close Relatives of the Median: Quartiles and Percentiles: describes why we might want to look at the 25th, 50th, and 75th percentiles when examining our data	101
GSS Example: TV Watching Rerun: shows how to find the quartiles for the frequency distribution we examined earlier in the chapter	101
Envisioning Variation: introduces the very important concept of variation, showing how it exists throughout our daily lives	102
Assessing Variation among Groups: covers how to find the range, the interquartile range, and, most importantly, the variance	103
GSS Example: Educational Attainment: offers an example of how the variance for education differs among racial groups	108
Visual Variation: The Shapes of Distributions: uses a hypothetical series of examples to show how we can observe different levels of variation graphically	109
Assessing the Variation of Individual Cases: introduces the standard deviation and z-scores, and shows how to calculate them and talk about them	111
GSS Example: Internet Usage among Racial Groups: illustrates the uniqueness of various types of people with regard to how many hours per week they use the Internet	113
Finding s and s^2 with a Frequency Distribution: explains how the formulas change once we're looking at the data in grouped form	115
Variation When There Is No Mean: introduces the Index of Qualitative Variation	116
GSS Example: Attitudes toward Government Spending: using the IQV, shows how variation in attitudes has changed over time for numerous social issues	123
Other Measures of Diversity: briefly discusses two common measures: the index of dissimilarity and the Gini coefficient	126
Literature Example: Cost of Medical Care: describes a student-led research project that exposed frightening levels of variation in medical costs	127
Conclusion	128

	SPSS Demonstration: shows students numerous ways to get descriptive statistics in SPSS, and how to represent them in graph form	128
	From Output to Presentation: models the typical way descriptive statistics are presented in professional settings	135
	Exercises: examine such topics as gender differences in attitudes toward feminists, variation in health for people in different income brackets, and teen texting habits	136
Chapter 4:	Using Sample Crosstabs to Talk about Populations: The Chi-Square Test	141
	Introduction: discusses the importance of sampling, the difference between populations and samples, and what an inference is	141
	A Series of Hypothetical Crosstabulations: illustrates different levels of relationship in a crosstab, from perfect to none, and how this relates to the inference we're trying to make based on these results	143
	Calculating Expected Frequencies: shows, based on what we've observed in our actual data, what these really mean (i.e., what are we actually expecting?)	146
	Introducing Chi-Square: goes through the calculation of the chi-square value step by step, using one of the hypothetical crosstabs we've already covered	151
	Statistical Significance: covers one of the most important concepts in social statistics, as it determines what we can say about our population based on our sample crosstab; explains, in reference to the hypothetical crosstabs, the idea of type one and type two error; shows how to use the chi-square table	153
	The Effect of Sample Size: drives home the point that the chi-square procedure and statistical significance are reliant on sample size	156
	Chi-Square with Larger Crosstabs: covers the concept of degrees of freedom and what, in the context of a crosstab, this actually means	157
	GSS Example: Gun Ownership and Sex: goes through the entire chi-square procedure again, this time using real data	160
	GSS Example: Age and Cynicism: emphasizes the importance of looking out for a few cells that are causing a high chi-square value and statistical significance	162
	The Language of Hypothesis Testing: introduces the oft-used language of null and alternative hypotheses	164
	The Chi-Square Distribution: shows, now that we know what chi-square does, what is really going on in terms of a probability	

	distribution; builds a chi-square distribution step by step to make clear the connection between this distribution and statistical significance	165
	GSS Example: Catholic Confidence in Organized Religion: combines chi-square with elaboration to show how time, gender, and confidence are related	171
	GSS Example: Guns, Age, and Sex: returns to an earlier example and shows how elaboration and chi-square can be used to summarize an interesting story	174
	Literature Example: Real Research about Poop, Really: goes over a fascinating example in which researchers used crosstabs and chi-square to illuminate differences based on sex and sexual orientation with regard to college students' paranoia over public restrooms	175
	Literature Example: Obesity in the United States and France: summarizes a cross-cultural study of media coverage of obesity, using crosstabs and chi-square to show differences in how this social problem is framed in the two countries	178
	Conclusion	180
	SPSS Demonstration: goes through how to conduct a chi-square test in SPSS and interpret the output	181
	From Output to Presentation: models how to present chi-square results in a professional format	184
	Exercises: examine such topics as the relationship between smoking and sexual orientation, the phenomenon of customer showrooming in retail stores, and parental propensity for friending their children on social networks	185
Chapter 5:	Using a Sample Mean or Proportion to Talk about a Population: Confidence Intervals	189
	Introduction: makes connections between what we did in the chi-square chapter and where we are headed in this chapter	189
	Sampling Distributions of Sample Means: illustrates the step-by-step building of a sampling distribution, showing the probabilities of pulling various sample means	190
	The Standard Error: makes very clear, based on the previous example, what the standard error really is, why it might be useful, and how we can estimate it	200
	Claims about the Population Mean: introduces the first use of these new ideas: using sample data to refute a claim someone makes about a population mean	203
	GSS Example: TV Watching among Young Women: puts to use the population claim procedure using real data	207

	Confidence Intervals: introduces the second use of these new ideas: based on sample data, confidently predicting where the population mean can fall	207
	GSS Example: Police Violence: builds confidence intervals for various racial groups with regard to their support for police use of violence	212
	GSS Example: Job Stress and Satisfaction: shows how variation in a sample can affect confidence interval width	215
	Confidence Intervals with Proportions: modifies the confidence interval procedure to build margins of error around sample proportions	215
	GSS Example: Support for Marijuana Legalization: puts the student in the situation of working for a politician speaking on this issue before two different age groups	218
	Literature Example: Black Characters in Film: recounts a content analysis of recent popular films that feature certain types of black characters, and uses confidence intervals to do it	218
	Conclusion	220
	SPSS Demonstrations: cover the important and often tricky procedure of selecting cases, as well as covering confidence intervals	220
	From Output to Presentation: shows how professionals often present confidence interval results	227
	Exercises: examine such topics as Internet news consumption, Hispanics' acquisition of health insurance, and teenagers engaging in "sexting"	228
Chapter 6:	Using Multiple Sample Means to Talk about Populations: <i>t</i>-Tests and ANOVA	231
	Introduction: makes connections between the previous two chapters and where this chapter will take the student	231
	A Different Kind of Sampling Distribution: builds, step by step, a sampling distribution of sample mean differences to illustrate the key concept of the chapter	232
	Testing Differences between Two Means: The <i>t</i> -test: introduces the essence of the <i>t</i> -test and how the language of hypothesis testing is used with the test	235
	GSS Example: Back to TV Watching: goes through the entire <i>t</i> -test procedure using real data	235
	Looking More Closely at the Formula: manipulates the results from the previous example to illustrate how the <i>t</i> -test formula really works	237

	GSS Example: Suicide, Age, and Political Party: compares strong Democrats and strong Republicans for three age groups on an index of suicide support	238
	Testing the Difference among More than Two Means: ANOVA: walks students through the ANOVA procedure, pointing out its similarities to the <i>t</i> -test and chi-square test	240
	A Comparative Graphical Approach to Understanding ANOVA: uses a series of three ANOVAs to show how variation affects the outcome	245
	GSS Example: Attitude versus Behavior about Housework: uses two ANOVAs to see if men's and women's attitudes affect their actual housework behavior	247
	Interchapter Connection: ANOVA and Chi-Square: describes how these two related tests both set up hypothetical situations to which to compare the actual data	250
	GSS Example: Internet Use, Race, and Gender: shows that, similar to the chi-square test, a single group can be responsible for a statistically significant ANOVA	250
	Literature Example: Overdoing Gender: recounts an article that uses <i>t</i> -tests to show that when men's gender is threatened, they react in very interesting ways	251
	Literature Example: Activism through the Life Course: covers an article that compares three groups of people who had different levels of activism in college in order to see how active they are a generation later	253
	Conclusion . . . with Interchapter Connections: reviews how to choose which technique to use based on what types of variables you have	255
	SPSS Demonstrations: covers <i>t</i> -tests and ANOVA	255
	From Output to Presentation: models how to present a series of <i>t</i> -test results	259
	Exercises: addresses such questions as: do those who vote trust government more than those who don't vote? Do women use technology to shop more than men? Does residential location affect health?	260
Chapter 7:	Give Me One Good Reason Why: Bivariate Correlation and Regression	264
	Introduction: starts the hypothetical example that runs through the chapter: the relationship between men's heights and incomes	264
	Linear Equations: reviews the basics of linear equations, such as constants and slopes	265

Calculating the Regression Equation: takes things step by step to calculate the slope and constant	267
Calculating the Correlation Coefficient: describes how to find the correlation coefficient and what this number means	273
The Effects of an Outlier: modifies one of the original data points from the previous example to show that, particularly in a small dataset, an outlier can have a huge effect on the regression results	276
Explained Variation: covers one of the most important concepts in regression, and does so in a way that explains exactly what explained variation is	277
Example: Forecasting with Regression: shows how regression can be used to forecast into the future, using data regarding movie grosses	281
GSS Example: Education and Income: uses real data to analyze a classic relationship using simple regression	285
GSS Example: Income and Hours to Relax: switches income to independent variable status to see if it affects how much time people have to relax, emphasizing that sometimes a lack of effect is the most interesting finding of all	288
GSS Example: Explaining Cynicism: uses several demographic variables, one at a time, of course, to explain why some people are more cynical than others	288
GSS Example: Intergenerational Family Size: runs the same regression for several groups, foreshadowing a more advanced technique later in the book	289
Literature Example: Support for the War on Terror: covers an article that uses correlations to examine relationships between authoritarianism and support for various civil-rights-restricting policies	290
Literature Example: Physical Attractiveness: recounts an article that uses correlations to examine, for both men and women, which factors are related to people's ratings of others' physical attractiveness	291
Conclusion	293
SPSS Demonstration: Creating a Scatterplot: shows how to create a visual representation of the relationship between two ratio-level variables	293
Exercises: analyze such topics as traditionalism, environmentalism, and AIDS	295

Chapter 8:	Using Sample Slopes to Talk about Populations: Inference and Regression	303
	Introduction: links regression-related inference with the types of inference already covered in the book	303
	One More Sampling Distribution: builds one last sampling distribution, showing how the repeated random sampling of slopes gives us another useful probability distribution	304
	From Standard Error to a <i>t</i>-Value to a <i>p</i>-Conclusion: introduces the standard error of the slope, and how to use it to conduct a <i>t</i> -test to determine the statistical significance of our sample slope, making connections to previous similar techniques, and emphasizing the importance of sample size	306
	GSS Example: Education and Sexual Activity: investigates the relationship between education and amount of sex, looking at different age groups with differing sample sizes	311
	GSS Example: Income and Sexual Activity: drives home the point that statistical significance can differ from substantive significance	313
	GSS Example: Work Time and Sexual Activity: provides another example how a lack of a statistically significant finding can still be thought of as interesting	314
	Literature Example: Grade Point Averages: recounts an article that tried to find evidence for a commonly assumed relationship: how much a student studies and her G.P.A.	314
	Literature Example: Family Size and Grades: tells the story of an article that investigates whether kids from large families have lower grades, and provides another distinction between statistical and substantive significance	316
	Conclusion	317
	SPSS Demonstration: Regression: walks through the SPSS Regression procedure and how to make sense of the output	317
	Creating a Correlation Matrix: shows how to run and read a set of correlations in the form of a matrix	320
	From Output to Presentation: presents the first example of what will soon become a theme: how to present regression results in a format similar to that in professional research	321
	Exercises: investigate such topics as the relationship between age and concern about privacy, the relationships among health conditions, and kids' perceptions of their parents' cyberblocking behavior	322

Chapter 9:	It's All Relative: Dichotomies as Independent Variables in Regression	326
	Introduction: recalls the importance of levels of measurement and why we may want to use variables other than those at the ratio level in regression	326
	Dichotomies: shows how, once you recode a dichotomy into 0 and 1, interpreting the slope of a dichotomy is quite simple	327
	Interchapter Connection: <i>t</i> -Tests versus Regression: reruns the previous regression example as a <i>t</i> -test, showing the similarities and differences between <i>t</i> -tests and regression	328
	Categorical Variables: introduces the technique of reference grouping, where we create dichotomies for all but one of the categories of a (typically) nominal-level variable, leaving one group as the reference category	329
	GSS Example: Variation in STEM Achievement: investigates variation in science and math achievement using sex, race, and political views as independent variables	331
	GSS Example: TV Watching and Work Status: shows how to use various work status variables as a set of reference-group variables	333
	GSS Example: Happiness and Partnership Status: investigates variation in happiness based on a set of reference-group variables regarding partnership status	334
	GSS Example: Party Identification and Political Knowledge: uses a set of reference-group variables to see if strength of political affiliation affects self-reported political knowledge	335
	Literature Example: Gender and Housework: examines an article that uses ratio-level variables, dichotomies, and reference groups to explain variation in time spent on housework	337
	Literature Example: Tracking Changes over Time: reports from an article that uses reference groups to show the changes over time in attitudes toward homosexuality	339
	Conclusion	341
	SPSS Demonstration: Reference Grouping: walks through the process of creating a set of reference-group variables and running a regression with them	341
	From Output to Presentation: continues the reporting of regression results in tabular form, showing how it is really useful when using reference groups	343
	Exercises: examine such issues as political efficacy, conspiracy theories, and economic optimism	344

Chapter 10: Above and Beyond: The Logic of Controlling and the Power of Nested Regression Models	348
Introduction: introduces the very important concept of statistical control by explaining how this concept exists throughout many aspects of our lives	348
GSS Example: Gender and Income: walks very step by step through a simple example, examining the effect of gender on income, and then how the effect diminishes but does not go away once we control for hours worked	350
A Different Way to Present Results: illustrates how to present results from multiple regression models side by side to emphasize how the models are nested inside of one another	352
GSS Example: Age and Income: takes students through an example very similar to the first GSS example, but this time using age instead of gender, showing how the effect of age completely disappears once we control for hours worked	354
GSS Example: Perceptions of U.S. Racial Makeup: offers another example of an original effect going away completely once we control for another variable of interest	355
Interchapter Connection: Controlling with Crosstabs: re-creates the previous example using hypothetical elaborated crosstabs to show how elaboration and nested models are related	358
GSS Example: Attitudes toward State Assistance: uses nested models to examine racial differences in support for welfare-related policies	362
GSS Example: Support for Same-Sex Parenting: this example introduces—after reviewing the various nested stories covered thus far—the type of nested story in which the original effect of interest grows upon the introduction of a new variable	364
Judging Improvement from Model to Model: briefly covers the <i>F</i> -test used in nested modeling to show if adding a new independent variable actually improved our ability to explain variation in the dependent variable	368
Sample Size from Model to Model: emphasizes the importance of keeping sample size constant from model to model, and gives an example of the consequences of not doing so	369
Literature Example: Oppositional Culture in Schools: explains how researchers used nested modeling to refute a popular explanation of racial differences in grades	370
Literature Example: Media Exposure and Fear of Crime: shows how a researcher used nested models to examine the commonly	

held assumption that certain types of media consumption lead people to be more fearful of crime	372
Conclusion	374
SPSS Demonstration: covers how to run a set of nested models, keeping the number of cases constant	375
From Output to Presentation: reviews how to present models side by side, and offers several tips for creating a clear table	377
Exercises: examine such topics as the Affordable Care Act, affirmative action, and the relationship between age of parents and the extent to which they monitor their children's technology use	378
Chapter 11: Some Slopes Are Bigger than Others: Calculating and Interpreting Beta Coefficients	384
Introduction: explains why comparing regular regression slopes is a very bad idea	384
The Process of Standardizing Slopes: takes a step-by-step approach to standardizing the slopes	385
A Shortcut: reveals, now that a full understanding of standardizing slopes has been reached, a quick shortcut to getting the betas	389
Interchapter Connection: Standardization and z-Scores: returns to z-scores and covers a regression example using z-scores to see what the betas look like	389
GSS Example: Religion and Abortion Attitudes: determines, through use of betas, which aspect of religiosity affects abortion attitudes the most	390
GSS Example: Following in Your Parents' Educational Footsteps: shows the gendered nature of educational generational connections	392
GSS Example: Gender and Happiness: examines how different aspects of life satisfaction (family, job, health) affect overall happiness for men and women	393
Literature Example: Racial Threat and School Discipline: covers a piece of research that uses betas in its regressions to examine racial attitudes and how they affect feelings about punishment	394
Literature Example: Country Music and Suicide: recounts a classic piece of social research that uses betas to show a very interesting relationship	395
Conclusion	397
SPSS Demonstration: reviews where in regression output one can find the betas	397
From Output to Presentation: models how to present both unstandardized and standardized slopes in the same table	398

	Exercises: explore such issues as the relationship between political views and attitudes toward business, which factors most affect the seeking of health information on the Internet, and ownership of technology	399
Chapter 12:	Different Slopes for Different Folks: Interaction Effects	402
	Introduction: describes the type of situation in which we would want to look at an interaction effect in contrast to other, simpler effects	402
	Interchapter Connection: Elaborated Crosstabs: eases our way into interaction effects by showing how they are connected to the elaborated crosstabs covered earlier in the book	404
	Creating the Interaction Effect: covers how to compute a new variable that multiplies two parent variables, at first with simply one ratio-level variable and one dichotomy (coded 0 and 1)	406
	GSS Example: Sex, Number of Children, and Relaxation: models how to tell the story of an interaction effect through the use of graphs	407
	GSS Example: Work Hours and Job Satisfaction: addresses the question: are men's and women's job satisfaction differentially affected by lengthy work hours?	410
	GSS Example: Civil Rights and Race: uses a set of controversial variables to see if whites and nonwhites are differentially affected by education.	412
	GSS Example: Race, Sex, and Religion: explores a different kind of interaction, one between two dichotomies	415
	GSS Example: Age, Education, and Sexual Activity: explores yet another kind of interaction, one between two ratio-level variables	416
	GSS Example: Knowing Someone with AIDS: goes back to 1988 GSS data to examine how knowing someone with AIDS interacts with political views	418
	Literature Example: Religion and Political Participation: plays around with results from an article that examines African American men and women and how their political activism is differentially affected by religious participation	420
	Literature Example: Gender, Work, and Guilt: returns to the example that started the chapter by working with results from an article about how men and women are differentially affected by work/home conflict	422
	Conclusion	425
	SPSS Demonstration: shows how to create an interaction effect and put it into a regression model	425

	From Output to Presentation: switches to Excel to show how to create professional graphs that illustrate an interaction effect	426
	Exercises: address such interaction-ready questions as: does income affect attitudes toward equality the same for whites and blacks? Does education have the same effect on technology acquisition for men as it does for women? Does saving money lead to happiness in all types of countries?	430
Chapter 13:	Explaining Dichotomous Outcomes: Logistic Regression	435
	Introduction: recaps what our dependent variables in regression have been thus far, and introduces the use of dichotomous dependent variables	435
	Regular Regression with a Dichotomous Dependent Variable: What Could Possibly Go Wrong? shows that bad things can happen when we do this	436
	What Logistic Regression Does: explains, on a fairly straightforward level, how logistic regression uses a natural logarithm and how that natural logarithm works	437
	GSS Example: Home Ownership: goes through the entire interpretive process of creating an equation, calculating a z-value, and using a simple little equation to change this z-value into a predicted probability	438
	GSS Example: Support of Gun Control: taking into account that the primary GSS question regarding gun control is asked dichotomously, this example shows how sex, age, and owning a gun affect gun control support	440
	GSS Example: Interracial Friendships: raises the possibility that, even if a variable is measured at the ratio level, its distribution might lead you to dichotomize it and use logistic regression	442
	GSS Example: Charitable Giving: combines logistic regression with nested modeling to explain why some people give to charity while others do not	445
	GSS Example: Capital Punishment: combines logistic regression with an interaction effect to explain the relationships among race, education, and support for capital punishment	447
	GSS Example: Condom Usage: uses another interaction effect to show how amount of sex and number of partners affects the dichotomous outcome of condom usage	449
	Another Way of Looking at Things: Odds Ratios: takes into account the fact that many social researchers use odds ratios to present their logistic regression results, and carefully covers how to interpret these ratios	450

	GSS Example: Capital Punishment Revisited: codes a dichotomous race variable in two different ways, showing how this affects the logistic results and odds ratios	453
	Literature Example: War and Presidential Disapproval: recounts a piece of research that uses a dichotomous measure of presidential disapproval to see if connections to 9/11 fatalities or war fatalities have an effect	455
	Literature Example: Global Warming: walks through an article that uses logistic regression and two interaction effects to explain attitudes toward global warming	456
	Conclusion	458
	SPSS Demonstration: Running a Logistic Regression: shows how to sift through the copious SPSS output from running a logistic regression	458
	From Output to Presentation: switches to Excel to show how to create a graph to convey logistic regression results	460
	Exercises: explore such topics as smoking, passport ownership, and having health insurance	462
Chapter 14:	Visualizing Causal Stories: Path Analysis	467
	Introduction: reviews how we've examined relationships up to this point, and previews how path analysis adds to our understanding	467
	GSS Example: Housework: carefully covers the details of path analysis, emphasizing the importance of direct and indirect effects	468
	Interchapter Connection: Nested versus Path: shows how path analysis has some similarities to nested modeling	473
	GSS Example: Same-Sex Parenting Revisited: revises an example from earlier in the book by using path analysis	474
	GSS Example: Education, Income, and Political Party: uses path analysis to analyze the sometimes contradictory relationships among these variables	476
	GSS Example: Explaining Drinking Behavior: investigates how education and age affect drinking behavior through indirect effects with bar going	478
	GSS Example: Like Father, Like Son? applies a classic use of path analysis: showing intergenerational status attainment relationships	481
	Literature Example: The Effects of Activism: introduces a classic example of the use of path analysis: to study how participating in high-risk activism affected the activists' subsequent lives	483

	Literature Example: Emotions in Service Work: recounts a study done in a bank, which used path analysis to see if positive emotions are contagious	485
	Conclusion	487
	SPSS Demonstration: teaches how to trick SPSS when running the models necessary for a path analysis	487
	From Output to Presentation: shows how to present a path model professionally	489
	Exercises: cover such topics as the contradictory relationships among age, religious attendance, and abortion attitudes; education, health problems, and health-related Internet use; and age, health problems, and health-related Internet use	489
Chapter 15:	Questioning the Greatness of Straightness: Nonlinear Relationships	493
	Introduction: explains how some relationships are better modeled as nonlinear relationships, using the common education-and-income example, showing that income does not increase in a linear fashion for every year of education	493
	Modeling Nonlinear Relationships: walks students through the process of creating a squared term, using hypothetical, small-sample-size examples; graphically shows the relationship between the squared and non-squared effects	497
	GSS Example: Age and Income: offers a classic example, using real data now, of age's nonlinear relationship with income	507
	GSS Example: Income and Financial Satisfaction: introduces the concept of diminishing returns and how we can examine such a phenomenon using nonlinear regression, such as the fact that rising income eventually loses its ability to increase financial satisfaction	510
	GSS Example: Education and Income: revisits the example that started the chapter, using a squared term now	513
	GSS Example: Income and Political Party: shows how the relationship between these two variables is linear for one racial group but nonlinear for another racial group, akin to an interaction effect	515
	Using Logarithms to Straighten out a Relationship: briefly covers how to transform variables using logarithms, especially when you have variables with huge ranges	517
	Literature Example: Housework and Gender: shows how a researcher modeled the nonlinear effect that gendered occupations have on men's propensity to do only "masculine" types of housework	521

	Literature Example: Effectiveness of Congresswomen: tells the story of how political scientists found that the relationship between vote share and effectiveness is nonlinear	523
	Conclusion	525
	SPSS Demonstration: covers how to create a squared term and put it into a regression model	526
	From Output to Presentation: returns to Excel to show how to create a graph to represent a nonlinear relationship	527
	Exercises: investigate such topics as the relationship between age and economic peril, middle-aged people being “sandwiched” by care responsibilities for their children and their parents, and the global relationship between health and happiness	528
Chapter 16:	Problems and Prospects: Regression Diagnostics, Advanced Techniques, and Where to Go Now	532
	Introduction: explains how this chapter will cover a few advanced topics at an “awareness” level	532
	Potential Problem One: Outliers: gives advice for how to spot whether an outlier might be having an effect on regression results	533
	Potential Problem Two: Multicollinearity: introduces a common problem in regression and discusses how to recognize if it is really a problem or not	537
	Advanced Techniques Concerning Variables: goes beyond regular and logistic regression and introduces on a very basic level ordered logistic, multinomial, probit, tobit, negative binomial, and Poisson regression	542
	Advanced Techniques Concerning Samples: introduces multilevel modeling, which is used with the complex samples so common in today’s social research	544
	Other Advanced Techniques: introduces structural equation modeling and hazard modeling	545
	No, Really, We’re Done, Go Home!: suggests how to continue your exploration of social statistics on your own	548
	Exercises: present scenarios that might call for these advanced techniques	549
Appendix A:	<i>Variables and Indexes from the Datasets Used in the End-of-Chapter Exercises</i>	<i>A-1</i>
	Codebooks of the variables used in the SPSS demonstrations and end-of-chapter exercises; explanations of the dozens of indexes created from these variables	

<i>Appendix B:</i>	<i>86 Articles That Use Statistics in Less than Scary Ways</i>	<i>B-1</i>
	Descriptions of 86 fascinating but relatively straightforward articles from sociology, political science, criminology, public health, and business; plus, for each article, confusing things to watch out for, and, for each article, a few questions to think about	
<i>Appendix C:</i>	<i>Statistical Tables</i>	<i>C-1</i>
<i>Appendix D:</i>	<i>Answers to Selected End-of-Chapter Exercises</i>	<i>D-1</i>
	<i>Bibliography</i>	<i>R-1</i>
	<i>Glossary/Index</i>	<i>I-1</i>

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 from 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 that 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

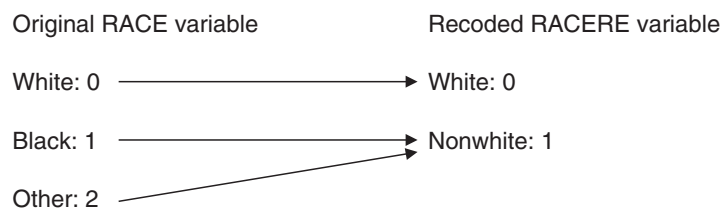
	SEX	AGE	RACE	DEGREE	CAPPUN
1	0	42	0	1	0
2	0	75	2	3	2
3	1	20	0	0	1
4	0	56	1	4	3
5	1	33	0	3	.
6	1	63	1	1	3

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

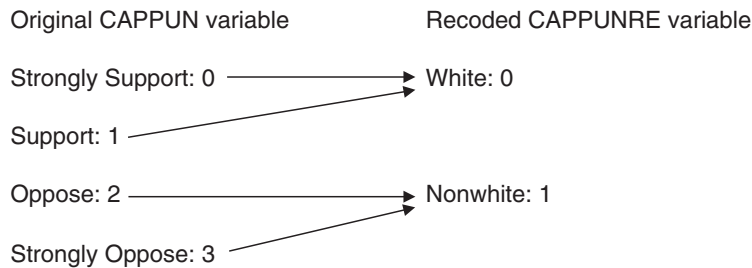
Most of the time, the data aren't exactly in the shape that we need them to be. But this is easy enough to fix. What we do is take the original variable and go through a process called **recoding** in order to create a new variable (always leaving the old variable in its original state). If we wanted 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