

STATISTICS USING STATA

An Integrative Approach

Second Edition

Building upon the success of the first edition, *Statistics Using Stata* uses the latest version of Stata to meet the needs of today's students. Engaging and accessible for students from a variety of mathematical backgrounds, this textbook integrates statistical concepts with the Stata (version 16) software package. It aligns Stata commands with examples based on real data, enabling students to understand statistics in a way that reflects statistical practice. Capitalizing on Stata's menu-driven "point and click" and program syntax interface, the chapters guide students from the comfortable "point and click" environment to the beginnings of statistical programming. Its coverage of essential topics gives instructors flexibility in curriculum planning and provides students with more advanced material to prepare for future work. Online resources – including solutions to exercises, PowerPoint slides, and Stata syntax (.do-files) for each chapter – allow students to review independently and adapt code to analyze new problems.

Sharon Lawner Weinberg is Professor of Applied Statistics and Psychology and former Vice Provost for Faculty Affairs at New York University, USA.

Sarah Knapp Abramowitz is Professor of Mathematics and Computer Science at Drew University, USA, where she is also Department Chair and Coordinator of Statistics Instruction.

Statistics Using Stata

AN INTEGRATIVE APPROACH

Second Edition

SHARON LAWNER WEINBERG

New York University

SARAH KNAPP ABRAMOWITZ

Drew University



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India
79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781108725835

DOI: 10.1017/9781108770163

First edition © Sharon Lawner Weinberg and Sarah Knapp Abramowitz 2016

Second edition © Sharon Lawner Weinberg and Sarah Knapp Abramowitz 2020

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First edition 2016

Second edition 2020

Printed in the United Kingdom by TJI International Ltd, Padstow Cornwall, 2020

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Weinberg, Sharon L., author. | Abramowitz, Sarah Knapp, 1967– author.

Title: Statistics using Stata : an integrative approach / Sharon Lawner

Weinberg, New York University, Sarah Knapp Abramowitz, Drew University.

Description: Cambridge, United Kingdom ; New York, NY, USA : University

Printing House, 2019. | Includes bibliographical references and index.

Identifiers: LCCN 2019019437 | ISBN 9781108725835 (alk. paper)

Subjects: LCSH: Mathematical statistics – Data processing. | Stata.

Classification: LCC QA276.4 .W455 2019 | DDC 519.50285/555–dc23

LC record available at <https://lcn.loc.gov/2019019437>

ISBN 978-1-108-72583-5 Paperback

Additional resources for this publication at www.cambridge.org/Stats-Stata2e

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

To our families

Contents

<i>Preface</i>	<i>page</i>	xv
New to the Second Edition		xv
Guiding Principles Underlying Our Approach		xvi
Overview of Content Coverage and Intended Audience		xvii
<i>Acknowledgments</i>		xix
1 INTRODUCTION		1
The Role of Statistical Software in Data Analysis		1
Statistics: Descriptive and Inferential		2
Variables and Constants		3
The Measurement of Variables		3
Nominal Level		4
Ordinal Level		4
Interval Level		5
Ratio Level		6
Choosing a Scale of Measurement		6
Discrete and Continuous Variables		8
Setting a Context with Real Data		11
Exercises		14
2 EXAMINING UNIVARIATE DISTRIBUTIONS		26
Counting the Occurrence of Data Values		26
When Variables are Measured at the Nominal Level		26
Frequency and Percent Distribution Tables		27
Bar Charts		28
Pie Charts		31
When Variables are Measured at the Ordinal, Interval, or Ratio Level		32
Frequency and Percent Distribution Tables		32
Stem-and-Leaf Displays		35
Histograms		38
Line Graphs		40
Describing the Shape of a Distribution		42
Accumulating Data		44
Cumulative Percent Distributions		44
Ogive Curves		45
Percentile Ranks		46
Percentiles		47
Five-Number Summaries and Boxplots		51
Modifying the Appearance of Graphs		56

	Summary of Graphical Selection	56
	Summary of Stata Commands	56
	Exercises	58
3	MEASURES OF LOCATION, SPREAD, AND SKEWNESS	74
	Characterizing the Location of a Distribution	74
	The Mode	74
	The Median	78
	The Arithmetic Mean	80
	<i>Interpreting the Mean of a Dichotomous Variable</i>	82
	<i>The Weighted Mean</i>	83
	Comparing the Mode, Median, and Mean	84
	Characterizing the Spread of a Distribution	86
	The Range and Interquartile Range	89
	The Variance	91
	The Standard Deviation	93
	Characterizing the Skewness of a Distribution	95
	Selecting Measures of Location and Spread	99
	Applying What We Have Learned	99
	Summary of Stata Commands	104
	Helpful Hints When Using Stata	105
	<i>Online Resources</i>	106
	<i>The Stata Command</i>	106
	<i>Stata Tips</i>	108
	Exercises	109
4	RE-EXPRESSING VARIABLES	118
	Linear and Nonlinear Transformations	118
	Linear Transformations: Addition, Subtraction, Multiplication, and Division	119
	The Effect on the Shape of a Distribution	121
	The Effect on Summary Statistics of a Distribution	121
	Common Linear Transformations	124
	Standard Scores	126
	z-Scores	127
	<i>Using z-Scores to Detect Outliers</i>	130
	<i>Using z-Scores to Compare Scores in Different Distributions</i>	133
	<i>Relating z-Scores to Percentile Ranks</i>	134
	Nonlinear Transformations: Square Roots and Logarithms	135
	Nonlinear Transformations: Ranking Variables	142
	Other Transformations: Recoding and Combining Variables	144
	Recoding Variables	144
	Combining Variables	147
	Data Management Fundamentals: The Do-File	147
	Summary of Stata Commands	150
	Exercises	151
5	EXPLORING RELATIONSHIPS BETWEEN TWO VARIABLES	159
	When Both Variables are at Least Interval-Leveled	159
	Scatterplots	160

CONTENTS

ix

	The Pearson Product-Moment Correlation Coefficient	166
	Interpreting the Pearson Correlation Coefficient	170
	<i>Judging the Strength of the Linear Relationship</i>	170
	<i>The Correlation Scale Itself Is Ordinal</i>	171
	<i>Correlation Does Not Imply Causation</i>	172
	<i>The Effect of Linear Transformations</i>	172
	<i>Restriction of Range</i>	173
	<i>The Shape of the Underlying Distributions</i>	174
	<i>The Reliability of the Data</i>	174
	When at Least One Variable Is Ordinal and the Other Is at Least Ordinal:	
	The Spearman Rank Correlation Coefficient	174
	When at Least One Variable Is Dichotomous: Other Special Cases of the	
	Pearson Correlation Coefficient	176
	The Point Biserial Correlation Coefficient: The Case of One at Least	
	Interval and One Dichotomous Variable	176
	The Phi Coefficient: The Case of Two Dichotomous Variables	181
	Other Visual Displays of Bivariate Relationships	185
	Selection of Appropriate Statistic or Graph to Summarize a Relationship	188
	Summary of Stata Commands	189
	Exercises	189
6	SIMPLE LINEAR REGRESSION	202
	The “Best-Fitting” Linear Equation	202
	The Accuracy of Prediction Using the Linear Regression Model	209
	The Standardized Regression Equation	210
	R As a Measure of the Overall Fit of the Linear Regression Model	210
	Simple Linear Regression When the Independent Variable Is	
	Dichotomous	214
	Using r and R As Measures of Effect Size	217
	Emphasizing the Importance of the Scatterplot	217
	Summary of Stata Commands	219
	Exercises	219
7	PROBABILITY FUNDAMENTALS	228
	The Discrete Case	228
	The Complement Rule of Probability	230
	The Additive Rules of Probability	231
	First Additive Rule of Probability	231
	Second Additive Rule of Probability	232
	The Multiplicative Rule of Probability	233
	The Relationship between Independence and Mutual Exclusivity	236
	Conditional Probability	236
	The Law of Total Probability	239
	Bayes’ Theorem	239
	The Law of Large Numbers	240
	Exercises	240
8	THEORETICAL PROBABILITY MODELS	244
	The Binomial Probability Model and Distribution	244
	The Applicability of the Binomial Probability Model	249

	The Normal Probability Model and Distribution	254
	Using the Normal Distribution to Approximate the Binomial Distribution	260
	Summary of Stata Commands	260
	Exercises	261
9	THE ROLE OF SAMPLING IN INFERENCE STATISTICS	269
	Samples and Populations	269
	Random Samples	270
	Obtaining a Simple Random Sample	271
	Sampling with and without Replacement	273
	Sampling Distributions	275
	Describing the Sampling Distribution of Means Empirically	275
	Describing the Sampling Distribution of Means Theoretically	280
	The Central Limit Theorem	281
	Estimators and Bias	285
	Summary of Stata Commands	286
	Exercises	287
10	INFERENCE INVOLVING THE MEAN OF A SINGLE POPULATION WHEN σ IS KNOWN	291
	Estimating the Population Mean, μ , When the Population Standard Deviation, σ , Is Known	291
	Interval Estimation	293
	Relating the Length of a Confidence Interval, the Level of Confidence, and the Sample Size	296
	Hypothesis Testing	296
	The Relationship between Hypothesis Testing and Interval Estimation	305
	Effect Size	306
	Type II Error and the Concept of Power	307
	Increasing the Level of Significance, α	310
	Increasing the Effect Size, δ	310
	Decreasing the Standard Error of the Mean, $\sigma_{\bar{x}}$	311
	Closing Remarks	312
	Summary of Stata Commands	313
	Exercises	314
11	INFERENCE INVOLVING THE MEAN WHEN σ IS NOT KNOWN: ONE- AND TWO-SAMPLE DESIGNS	319
	Single Sample Designs When the Parameter of Interest Is the Mean and σ Is Not Known	319
	The t -Distribution	320
	Degrees of Freedom for the One-Sample t -Test	321
	Violating the Assumption of a Normally Distributed Parent Population in the One-Sample t -Test	322
	Confidence Intervals for the One-Sample t -Test	323
	Hypothesis Tests: The One-Sample t -Test	330
	Effect Size for the One-Sample t -Test	333
	Two-Sample Designs When the Parameter of Interest Is μ , and σ Is Not Known	336
	Independent (or Unrelated) and Dependent (or Related) Samples	337

CONTENTS

xi

Independent Samples <i>t</i> -Test and Confidence Interval	338
The Assumptions of the Independent Samples <i>t</i> -Test	340
Effect Size for the Independent Samples <i>t</i> -Test	349
Paired Samples <i>t</i> -Test and Confidence Interval	353
The Assumptions of the Paired Samples <i>t</i> -Test	354
Effect Size for the Paired Samples <i>t</i> -Test	359
The Bootstrap	360
Conducting Power Analyses for <i>t</i>-Tests on Means	364
Summary	369
Summary of Stata Commands	372
Exercises	374
12 RESEARCH DESIGN: INTRODUCTION AND OVERVIEW	391
Questions and their Link to Descriptive, Relational, and Causal Research Studies	391
The Need for a Good Measure of our Construct: Weight	391
The Descriptive Study	392
From Descriptive to Relational Studies	393
From Relational to Causal Studies	393
The Gold Standard of Causal Studies: The True Experiment and Random Assignment	395
Comparing Two Kidney Stone Treatments Using a Non-Randomized Controlled Study	396
Including Blocking in a Research Design	397
Underscoring the Importance of Having a True Control Group Using Randomization	398
Analytic Methods for Bolstering Claims of Causality from Observational Data	402
Quasi-Experimental Designs	404
Threats to the Internal Validity of a Quasi-Experimental Design	404
Threats to the External Validity of a Quasi-Experimental Design	405
Threats to the Validity of a Study: Some Clarifications and Caveats	406
Threats to the Validity of a Study: Some Examples	407
Exercises	408
13 ONE-WAY ANALYSIS OF VARIANCE	412
The Disadvantage of Multiple <i>t</i>-Tests	412
The One-Way Analysis of Variance	414
A Graphical Illustration of the Role of Variance in Tests on Means	414
ANOVA As an Extension of the Independent Samples <i>t</i> -Test	416
Developing an Index of Separation for the Analysis of Variance	416
Carrying Out the ANOVA Computation	417
<i>The Between Group Variance (MS_B)</i>	418
<i>The Within Group Variance (MS_W)</i>	418
The Assumptions of the One-Way ANOVA	419
Testing the Equality of Population Means: The <i>F</i> -Ratio	420
How to Read the Tables and Use Stata Functions for the <i>F</i> -Distribution	422
ANOVA Summary Table	425
Measuring the Effect Size	426
Post-Hoc Multiple Comparison Tests	431

	The Bonferroni Adjustment: Testing Planned Comparisons	444
	The Bonferroni Tests on Multiple Measures	446
	Conducting Power Analyses for One-Way ANOVA	447
	Summary of Stata Commands	450
	Exercises	451
14	TWO-WAY ANALYSIS OF VARIANCE	457
	The Two-Factor Design	457
	The Concept of Interaction	460
	The Hypotheses That are Tested by a Two-Way Analysis of Variance	465
	Assumptions of the Two-Way Analysis of Variance	466
	Balanced versus Unbalanced Factorial Designs	467
	Partitioning the Total Sum of Squares	468
	Using the <i>F</i> -Ratio to Test the Effects in Two-Way ANOVA	469
	Carrying Out the Two-Way ANOVA Computation by Hand	469
	Decomposing Score Deviations about the Grand Mean	474
	Modeling Each Score As a Sum of Component Parts	475
	Explaining the Interaction As a Joint (or Multiplicative) Effect	475
	Measuring Effect Size	476
	Fixed versus Random Factors	479
	Post-Hoc Multiple Comparison Tests	479
	Simple Effects and Pairwise Comparisons	482
	Summary of Steps to Be Taken in a Two-Way ANOVA Procedure	487
	Conducting Power Analyses for Two-Way ANOVA	491
	Summary of Stata Commands	493
	Exercises	495
15	CORRELATION AND SIMPLE REGRESSION AS INFERENTIAL TECHNIQUES	503
	The Bivariate Normal Distribution	503
	Testing whether the Population Pearson Product-Moment Correlation Equals Zero	506
	Using a Confidence Interval to Estimate the Size of the Population Correlation Coefficient, ρ	509
	Revisiting Simple Linear Regression for Prediction	512
	Estimating the Population Standard Error of Prediction, $\sigma_{Y X}$	513
	Testing the <i>b</i> -Weight for Statistical Significance	514
	Explaining Simple Regression Using an Analysis of Variance Framework	518
	Measuring the Fit of the Overall Regression Equation: Using R and R^2	520
	Relating R^2 to $\sigma^2_{Y X}$	521
	Testing R^2 for Statistical Significance	522
	Estimating the True Population R^2 : The Adjusted R^2	523
	Exploring the Goodness of Fit of the Regression Equation: Using Regression Diagnostics	524
	Residual Plots: Evaluating the Assumptions Underlying Regression	526
	Detecting Influential Observations: Discrepancy and Leverage	529
	Using Stata to Obtain Leverage	530
	Using Stata to Obtain Discrepancy	531
	Using Stata to Obtain Influence	531
	Using Diagnostics to Evaluate the Ice Cream Sales Example	533

CONTENTS

xiii

Using the Prediction Model to Predict Ice Cream Sales	536
Simple Regression When the Predictor Is Dichotomous	536
Conducting Power Analyses for Correlation and Simple Regression	538
Summary of Stata Commands	540
Exercises	541
16 AN INTRODUCTION TO MULTIPLE REGRESSION	553
The Basic Equation with Two Predictors	554
Equations for b , β and $R_{Y.12}$ When the Predictors Are Not Correlated	555
Equations for b , β , and $R_{Y.12}$ When the Predictors Are Correlated	556
Summarizing and Expanding on Some Important Principles of Multiple Regression	558
Testing the b -Weights for Statistical Significance	563
Assessing the Relative Importance of the Predictors in the Equation	565
Measuring the Drop in R^2 Directly: An Alternative to the Squared Semipartial Correlation	566
Evaluating the Statistical Significance of the Change in R^2	566
The b -Weight As a Partial Slope in Multiple Regression	568
Multiple Regression When One of the Two Independent Variables Is Dichotomous	571
Controlling Variables Statistically: A Closer Look	576
A Hypothetical Example	577
Conducting Power Analyses for Multiple Regression	580
Summary of Stata Commands	582
Exercises	583
17 TWO-WAY INTERACTIONS IN MULTIPLE REGRESSION	590
Testing the Statistical Significance of an Interaction Using Stata	593
Comparing the \hat{Y} Values from the Additive and Interaction Models	598
Centering First-Order Effects if the Equation Has an Interaction	599
Probing the Nature of a Two-Way Interaction	600
Interaction When One of the Independent Variables Is Dichotomous and the Other Is Continuous	603
Methods Useful for Model Selection	610
Conducting a Power Analysis to Detect an Interaction	613
Summary of Stata Commands	614
Exercises	617
18 NONPARAMETRIC METHODS	622
Parametric versus Nonparametric Methods	622
Nonparametric Methods When the Dependent Variable Is at the Nominal Level	623
The Chi-Square Distribution (χ^2)	623
The Chi-Square Goodness-of-Fit Test	625
The Chi-Square Test of Independence	630
<i>Assumptions of the Chi-Square Test of Independence</i>	633
Fisher's Exact Test	635
<i>Calculating the Fisher's Exact Test by Hand Using the Hypergeometric Distribution</i>	637

	Nonparametric Methods When the Dependent Variable Is Ordinal-Leveled	639
	Wilcoxon Sign Test	640
	The Mann–Whitney <i>U</i> -Test or Wilcoxon’s Rank-Sum Test	642
	The Kruskal–Wallis Analysis of Variance	647
	Summary of Stata Commands	649
	Exercises	650
19	COMMUNICATING YOUR STATA RESULTS VIA EXCEL	655
	Setting the Working Directory	655
	Reproducing a Table of Univariate Summary Statistics in Excel	656
	Using estpost and esttab	656
	Using putexcel	657
	Reproducing a Correlation Matrix As a Table in Excel	661
	Using estpost and esttab	661
	Using putexcel	662
	Reproducing Regression Output As a Table in Excel	663
	Using outreg2 to obtain a table of model statistics in Excel	663
	Using eststo and esttab to obtain a table of model statistics in Excel	663
	Using putexcel to reproduce a table of regression coefficients in Excel	664
	Reproducing a Graph in Excel (Using putexcel)	666
	Conclusion	668
	Summary of Stata Commands	668
	Exercises	671
	<i>Appendix A Data Set Descriptions</i>	673
	<i>Appendix B Stata .Do-files and Data Sets in Stata Format</i>	686
	<i>Appendix C Statistical Tables</i>	688
	<i>Appendix D Solutions</i>	711
	<i>References</i>	712
	<i>Index</i>	716

Preface

This second edition of *Statistics Using Stata: An Integrative Approach* continues to capitalize on the versatility and power of the Stata software package to create a course of study that links good statistical and data science practice to the analysis of real data. It also benefits from the many years of the authors' experience teaching statistics to undergraduate students at a liberal arts university and to undergraduate and graduate students at a large research university from a variety of disciplines including education, psychology, health, and policy analysis. Stata provides both a menu-driven and command-line approach to the analysis of data and, in so doing, facilitates the smooth transition to a more advanced course of study in statistics.

New to the Second Edition

While the first edition of this text is based on Stata's version 14, the second edition is based on version 16, the most current version of Stata. We devote an entirely new chapter (Chapter 19) on the command **putexcel** new to version 15 and highlighted in version 16, so that readers may produce reproducible reports that include professionally styled tables and figures based on the statistical analyses they have learned from throughout the text. As we know, drafting a cohesive, thoughtful, and compelling narrative is critically important for enabling work to be heard and appreciated. Through multiple examples in this new chapter we illustrate how **putexcel** may be used along with other Stata commands (e.g., **estpost** and **esttab**) to substantiate that narrative with professionally styled tables and figures that are concise, cleanly illustrated, and easily reproducible.

In the first edition, we gave students an important conceptual understanding of power and its importance in designing a research study. In the second edition, we extend that understanding to a practical level by providing readers with knowledge about how to use Stata to carry out a power analysis for each of the inferential methods covered in the text. In particular, we illustrate using Stata's **power** command how to link the key elements of power analysis to the calculation of a minimum sample size needed to obtain a specified power to detect a statistically significant and meaningful result.

To give greater clarity to the notion of statistical control, in the second edition we expand upon the discussion of statistical control in our first edition and include examples that illustrate in what sense multiple regression allows for the statistical control of confounders. In the second edition, we elaborate upon the use of the adjusted R -squared as a measure of model goodness-of-fit to its use as an aid in model selection. We also include discussion of cross-validation and the AIC and BIC statistics as more modern

approaches to model selection that are frequently referenced in the data science and research literature.

Because students often have a difficult time understanding and interpreting statistical interaction within the multiple regression setting, in the second edition we devote an entirely new chapter to this topic. In this new Chapter 17, we emphasize the importance of visualization via the versatile **margins** and **marginsplot** commands, which are relatively new additions to Stata. We also include in this edition the testing of simple effects following a significant interaction. We do so as a way to help uncover and elucidate the nuanced nature of the interaction obtained. Finally, although we have made a substantial number of changes in the second edition from the first, the second edition, like its predecessor, embraces and is motivated by the following important guiding principles.

Guiding Principles Underlying Our Approach

First, and perhaps most importantly, we believe that a good data analytic plan must serve to uncover the story behind the numbers; what the data tell us about the phenomenon under study. To begin, a good data analyst must know his/her data well and have confidence that it satisfies the underlying assumptions of the statistical methods used. Accordingly, we emphasize the usefulness of diagnostics in both graphical and statistical form to expose anomalous cases, which might unduly influence results, and to help in the selection of appropriate assumption-satisfying transformations so that ultimately we may have confidence in our findings. We also emphasize the importance of using more than one method of analysis to answer fully the question posed and understanding potential bias in the estimation of population parameters.

Second, because we believe that data are central to the study of good statistical practice, the textbook's website contains several data sets used throughout the text. Two are large sets of real data that we make repeated use of in both worked-out examples and end-of-chapter exercises. One data set contains 48 variables and 500 cases from the education discipline that includes psychosocial measures such as self-concept; the other contains 49 variables and nearly 4,500 cases from the health discipline. By posing interesting questions about variables in these large, real data sets (e.g., Is there a gender difference in eighth graders' expected income at age 30? Is self-concept in eighth grade related to eighth graders' expected income at age 30?), we are able to employ a more meaningful and contextual approach to the introduction of statistical methods and to engage students more actively in the learning process. The repeated use of these data sets also contributes to creating a more cohesive presentation of statistics; one that links different methods of analysis to each other and avoids the perception that statistics is an often-confusing array of so many separate and distinct methods of analysis, with no bearing or relationship to one another.

Third, to facilitate the analysis of these data, and to provide students with a platform for actively engaging in the learning process associated with what it means to be a good researcher and data analyst, we incorporate the latest version of Stata (version 16) into the presentation of statistical material using a highly integrative approach that reflects practice. Students learn Stata along with each new statistical method covered, thereby allowing them to apply their newly learned knowledge to the real world of applications. In addition to demonstrating the use of Stata within each chapter, all chapters have an associated .do-file,

that may be accessed from the textbook's website (www.cambridge.org/Stats-Stata2e) along with other ancillary materials and resources. The .do-files are designed to allow students not only to replicate all worked out examples within a chapter, but also to reproduce the figures embedded in a chapter, and to create their own .do-files by extracting and modifying commands from them. Emphasizing data workflow management throughout the text using the Stata .do-file allows students to begin to appreciate one of the key ingredients in being a good researcher. Of course, another key ingredient to being a good researcher is content knowledge, and toward that end, we have included in the text a more comprehensive coverage of essential topics in statistics not covered by other textbooks at the introductory level, including robust methods of estimation based on resampling using the bootstrap, regression to the mean, the weighted mean, and potential sources of bias in the estimation of population parameters based on the analysis of data from quasi-experimental designs. We also include a discussion of Simpson's paradox, counterfactuals, and other issues related to research design in an entire chapter (Chapter 12) devoted to this topic.

Fourth, in accordance with our belief that the result of a null hypothesis test (to determine whether an effect is real or merely apparent) is only a means to an end (to determine whether the effect is important or useful) rather than an end in itself, we stress the need to evaluate the magnitude of an effect if it is deemed to be real, and of drawing clear distinctions between statistically significant and substantively significant results. Toward this end, we introduce the computation of standardized measures of effect size as common practice following a statistically significant result. While we provide guidelines for evaluating, in general, the magnitude of an effect, we encourage readers to think more subjectively about the magnitude of an effect, bringing into the evaluation their own knowledge and expertise in a particular area.

Finally, we believe that a key ingredient of an introductory statistics text is a lively, clear, conceptual, yet rigorous approach. We emphasize conceptual understanding through an exploration of both the mathematical principles underlying statistical methods and real-world applications. We use an easy-going, informal style of writing that we have found gives readers the impression that they are involved in a personal conversation with the authors. And we sequence concepts with concern for student readiness, reintroducing topics in a spiraling manner to provide reinforcement and promote the transfer of learning.

Overview of Content Coverage and Intended Audience

Along with the earlier topics mentioned, the inclusion of linear and nonlinear transformations, diagnostic tools for the analysis of model fit, tests of inference, an in-depth discussion of interaction and its interpretation in both two-way analysis of variance and multiple regression, and non-parametric statistics, the text provides a highly comprehensive coverage of essential topics in introductory statistics, and in so doing gives instructors flexibility in curriculum planning and provides students with more advanced material for future work in statistics. In addition, the text includes a large bibliography of references to relevant books and journal articles, and many end-of-chapter exercises with detailed answers on the textbook's website.

The book, consisting of 19 chapters, is intended for use in a one- or two-semester introductory applied statistics course for the behavioral, social, or health sciences at either the graduate or undergraduate level, or as a reference text as well. It is not intended for readers who wish to acquire a more theoretical understanding of mathematical statistics. To offer another perspective, the book may be described as one that begins with modern approaches to exploratory data analysis (EDA) and descriptive statistics, and then covers material similar to what is found in an introductory mathematical statistics text, designed, for example, for undergraduates in math and the physical sciences, but stripped of calculus and linear algebra, and grounded instead in data examples. Thus, theoretical probability distributions, Bayes' theorem, the law of large numbers, sampling distributions, and the central limit theorem are all covered, but in the context of solving practical and interesting problems.

Acknowledgments

This book has benefited from the many helpful comments of our New York University and Drew University students, too numerous to mention by name, and from the insights and suggestions of several colleagues. For their help, we would like to thank (in alphabetical order) Chris Apelian, Ellie Buteau, Chris Casement, Miao Chi, Sarah Friedman, Daphna Harel, Jennifer Hill, Chuck Huber, Michael Karchmer, Steve Kass, Jon Kettenring, Linda Lesniak, Yi Lu, Kathleen Madden, Isaac Maddow-Zimet, Meghan McCormick, Joel Middleton, and Marc Scott. Of course, any errors or shortcomings in the text remain the responsibility of the authors.