

STATISTICS USING R

An Integrative Approach

Using numerous examples with real data, this textbook closely integrates the learning of statistics with the learning of R. It is suitable for introductory-level learners, allows for curriculum flexibility, and includes, as an online resource, R-code script files for all examples and figures included in each chapter, for students to learn from and adapt and use in their future data analytic work. Other unique features created specifically for this textbook include an online R tutorial that introduces readers to data frames and other basic elements of the R architecture, and a CRAN library of datasets and functions that is used throughout the book. Essential topics often overlooked in other introductory texts, such as data management, are covered. The textbook includes online solutions to all end-of-chapter exercises and PowerPoint slides for all chapters as additional resources, and is suitable for those who do not have a strong background in mathematics.

Sharon Lawner Weinberg is Professor of Applied Statistics and Psychology, and the former Vice Provost for Faculty Affairs, at New York University (NYU). She is the recipient of the NYU Steinhardt Outstanding Teaching Award, and has taught statistics at both undergraduate and graduate levels. Her research has been supported by federal agencies and private foundations.

Daphna Harel is Associate Professor of Applied Statistics at New York University. She is known for her innovative approach to teaching both introductory and advanced statistics. Her research has been supported by federal agencies and foundations, such as the National Institutes for Health and the Canadian Institutes for Health Research.

Sarah Knapp Abramowitz is Professor of Mathematics and Computer Science, Department Chair, and Co-ordinator of Statistics Instruction at Drew University. She is Associate Editor of the *Journal of Statistics Education* and has presented at national conferences on topics related to the teaching of statistics.

Statistics Using R

AN INTEGRATIVE APPROACH

SHARON LAWNER WEINBERG

New York University

DAPHNA HAREL

New York University

SARAH KNAPP ABRAMOWITZ

Drew University, New Jersey



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/9781108719148
DOI: 10.1017/9781108755351
© Sharon Lawner Weinberg, Daphna Harel, and Sarah Knapp Abramowitz 2021
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 published 2021
A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Weinberg, Sharon L., author. | Harel, Daphna, 1988– author. |
Abramowitz, Sarah Knapp, 1967– author.
Title: Statistics using R : an integrative approach / Sharon Lawner
Weinberg, Daphna Harel, Sarah Knapp Abramowitz.
Description: Cambridge ; New York, NY : Cambridge University Press, 2021. |
Includes bibliographical references and index.
Identifiers: LCCN 2019059928 (print) | LCCN 2019059929 (ebook) |
ISBN 9781108719148 (paperback) | ISBN 9781108755351 (ebook)
Subjects: LCSH: R (Computer program language) |
Mathematical statistics – Data processing. | Statistics – Data processing.
Classification: LCC QA276.45.R3 W45 2021 (print) | LCC QA276.45.R3 (ebook) |
DDC 519.50285/5133–dc23
LC record available at <https://lcn.loc.gov/2019059928>
LC ebook record available at <https://lcn.loc.gov/2019059929>

ISBN 978-1-108-71914-8 Paperback
Additional resources for this publication at www.cambridge.org/stats-r

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.

Cambridge University Press
978-1-108-71914-8 — Statistics Using R
Sharon Lawner Weinberg , Daphna Harel , Sarah Knapp Abramowitz
Frontmatter
[More Information](#)

To our families

Contents

<i>Preface</i>	<i>page</i> xv
Guiding Principles Underlying Our Approach	xv
Overview of Content Coverage and Intended Audience	xvii
<i>Acknowledgments</i>	xviii
1 INTRODUCTION	1
The Role of Statistical Software in Data Analysis	2
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
Downloading and Interfacing with R	11
Installing and Loading R Packages	13
Setting a Context with Real Data	15
Exercises	17
2 EXAMINING UNIVARIATE DISTRIBUTIONS	25
Counting the Occurrence of Data Values	25
When Variables are Measured at the Nominal Level	25
Frequency and Percent Distribution Tables	26
Bar Charts	27
Pie Charts	29
When Variables are Measured at the Ordinal, Interval, or Ratio Level	30
Frequency and Percent Distribution Tables	30
Stem-and-Leaf Displays	33
Histograms	35
Line Graphs	37
Describing the Shape of a Distribution	41
Accumulating Data	42
Cumulative Percent Distributions	43
Percentile Ranks	44
Percentiles	45
Six-Number Summaries and Boxplots	48

	Summary of Graphical Selection	53
	Summary of R Commands	53
	Exercises	55
3	MEASURES OF LOCATION, SPREAD, AND SKEWNESS	70
	Characterizing the Location of a Distribution	70
	The Mode	70
	The Median	74
	The Arithmetic Mean	76
	<i>Interpreting the Mean of a Dichotomous Variable</i>	77
	<i>The Weighted Mean</i>	78
	Comparing the Mode, Median, and Mean	79
	Characterizing the Spread of a Distribution	82
	The Range and Interquartile Range	85
	The Variance	87
	The Standard Deviation	89
	Characterizing the Skewness of a Distribution	90
	Selecting Measures of Location and Spread	92
	Applying What We Have Learned	93
	Summary of R Commands	97
	Exercises	99
4	RE-EXPRESSING VARIABLES	108
	Linear and Nonlinear Transformations	108
	Linear Transformations: Addition, Subtraction, Multiplication, and Division	109
	The Effect on the Shape of a Distribution	111
	The Effect on Summary Statistics of a Distribution	112
	Common Linear Transformations	115
	Standard Scores	117
	z-Scores	118
	<i>Using z-Scores to Detect Outliers</i>	121
	<i>Using z-Scores to Compare Scores in Different Distributions</i>	123
	<i>Relating z-Scores to Percentile Ranks</i>	124
	Nonlinear Transformations: Square Roots and Logarithms	125
	Nonlinear Transformations: Ranking Variables	131
	Other Transformations: Recoding and Combining Variables	134
	Recoding Variables	134
	Combining Variables	135
	Data Management Fundamentals: The .R File	136
	Summary of R Commands	138
	Exercises	139
5	EXPLORING RELATIONSHIPS BETWEEN TWO VARIABLES	147
	When Both Variables are at Least Interval-Levelled	147
	Scatterplots	148
	The Pearson Product-Moment Correlation Coefficient	156
	Interpreting the Pearson Correlation Coefficient	160
	<i>Judging the Strength of the Linear Relationship</i>	160
	<i>The Correlation Scale Itself Is Ordinal</i>	161
	<i>Correlation Does Not Imply Causation</i>	162

CONTENTS

ix

	<i>The Effect of Linear Transformations</i>	162
	<i>Restriction of Range</i>	163
	<i>The Shape of the Underlying Distributions</i>	164
	<i>The Reliability of the Data</i>	164
	When at Least One Variable Is Ordinal and the Other Is at Least Ordinal:	
	The Spearman Rank Correlation Coefficient	164
	When at Least One Variable Is Dichotomous: Other Special Cases	
	of the Pearson Correlation Coefficient	166
	The Point Biserial Correlation Coefficient: The Case of One at Least	
	Interval and One Dichotomous Variable	166
	The Phi Coefficient: The Case of Two Dichotomous Variables	171
	Other Visual Displays of Bivariate Relationships	176
	Selection of Appropriate Statistic or Graph to Summarize a Relationship	180
	Summary of R Commands	181
	Exercises	182
6	SIMPLE LINEAR REGRESSION	193
	The “Best-Fitting” Linear Equation	193
	The Accuracy of Prediction Using the Linear Regression Model	201
	The Standardized Regression Equation	202
	R as a Measure of the Overall Fit of the Linear Regression Model	202
	Simple Linear Regression When the Predictor Variable Is Dichotomous	207
	Using <i>r</i> and <i>R</i> As Measures of Effect Size	209
	Emphasizing the Importance of the Scatterplot	209
	Summary of R Commands	212
	Exercises	212
7	PROBABILITY FUNDAMENTALS	221
	The Discrete Case	221
	The Complement Rule of Probability	223
	The Additive Rules of Probability	224
	First Additive Rule of Probability	224
	Second Additive Rule of Probability	225
	The Multiplicative Rule of Probability	227
	Conditional Probability	229
	The Relationship between Independence, Mutual Exclusivity, and	
	Conditional Probability	232
	The Law of Total Probability	233
	Bayes’ Theorem	234
	The Law of Large Numbers	236
	Summary of R Commands	237
	Exercises	237
8	THEORETICAL PROBABILITY MODELS	242
	The Binomial Probability Model and Distribution	242
	The Applicability of the Binomial Probability Model	247
	The Normal Probability Model and Distribution	251
	Using the Normal Distribution to Approximate the Binomial Distribution	259
	Summary of R Commands	259
	Exercises	260

9	THE ROLE OF SAMPLING IN INFERENTIAL STATISTICS	268
	Samples and Populations	268
	Random Samples	269
	Obtaining a Simple Random Sample	270
	Sampling with and without Replacement	271
	Sampling Distributions	273
	Describing the Sampling Distribution of Means Empirically	273
	Describing the Sampling Distribution of Means Theoretically	277
	The Central Limit Theorem	278
	Estimators and Bias	282
	Summary of R Commands	283
	Exercises	284
10	INFERENCES INVOLVING THE MEAN OF A SINGLE POPULATION WHEN σ IS KNOWN	287
	Estimating the Population Mean, μ , When the Population Standard Deviation, σ , Is Known	287
	Interval Estimation	289
	Relating the Length of a Confidence Interval, the Level of Confidence, and the Sample Size	293
	Hypothesis Testing	293
	The Relationship between Hypothesis Testing and Interval Estimation	302
	Effect Size	303
	Type II Error and the Concept of Power	304
	Increasing the Level of Significance, α	307
	Increasing the Effect Size, δ	308
	Decreasing the Standard Error of the Mean, $\sigma_{\bar{x}}$	308
	Closing Remarks	309
	Summary of R Commands	310
	Exercises	311
11	INFERENCES INVOLVING THE MEAN WHEN σ IS NOT KNOWN: ONE- AND TWO-SAMPLE DESIGNS	316
	Single Sample Designs When the Parameter of Interest Is the Mean and σ Is Not Known	316
	The t -Distribution	317
	Degrees of Freedom for the One-Sample t -Test	318
	Violating the Assumption of a Normally Distributed Parent Population in the One-Sample t -Test	319
	Confidence Intervals for the One-Sample t -Test	320
	Hypothesis Tests: The One-Sample t -Test	327
	Effect Size for the One-Sample t -Test	329
	Two-Sample Designs When the Parameter of Interest is μ , and σ Is Not Known	332
	Independent (or Unrelated) and Dependent (or Related) Samples	333
	Independent Samples t -Test and Confidence Interval	335
	The Assumptions of the Independent Samples t -Test	336
	Effect Size for the Independent Samples t -Test	344
	Paired Samples t -Test and Confidence Interval	346
	The Assumptions of the Paired Samples t -Test	347
	Effect Size for the Paired Samples t -Test	351

CONTENTS

xi

Conducting Power Analyses for <i>t</i> -Tests On Means	353
Summary	356
Summary of R Commands	360
Exercises	363
12 RESEARCH DESIGN: INTRODUCTION AND OVERVIEW	378
Questions and their Link to Descriptive, Relational, and Causal Research Studies	378
The Need for a Good Measure of our Construct: Weight	378
The Descriptive Study	379
From Descriptive to Relational Studies	380
From Relational to Causal Studies	380
The Gold Standard of Causal Studies: The True Experiment and Random Assignment	382
Comparing Two Kidney Stone Treatments Using a Non-Randomized Controlled Study	383
Including Blocking in a Research Design	384
Underscoring the Importance of Having a True Control Group Using Randomization	385
Analytic Methods for Bolstering Claims of Causality from Observational Data	389
Quasi-Experimental Designs	391
Threats to the Internal Validity of a Quasi-Experimental Design	392
Threats to the External Validity of a Quasi-Experimental Design	393
Threats to the Validity of a Study: Some Clarifications and Caveats	393
Threats to the Validity of a Study: Some Examples	394
Exercises	395
13 ONE-WAY ANALYSIS OF VARIANCE	399
The Disadvantage of Multiple <i>t</i> -Tests	399
The One-Way Analysis of Variance	401
A Graphical Illustration of the Role of Variance in Tests on Means	401
ANOVA As an Extension of the Independent Samples <i>t</i> -Test	402
Developing an Index of Separation for the Analysis of Variance	404
Carrying Out the ANOVA Computation	404
The Between Group Variance (<i>MS_B</i>)	405
The Within Group Variance (<i>MS_W</i>)	405
The Assumptions of the One-Way ANOVA	406
Testing the Equality of Population Means: The <i>F</i> -Ratio	407
How to Read the Tables and Use R Functions for the <i>F</i> -Distribution	408
The ANOVA Summary Table	413
Measuring the Effect Size	414
Post-Hoc Multiple Comparison Tests	416
The Bonferroni Adjustment: Testing Planned Comparisons	427
The Bonferroni Tests on Multiple Measures	429
Conducting Power Analyses for One-Way ANOVA	430
Summary of R Commands	431
Exercises	432

14	TWO-WAY ANALYSIS OF VARIANCE	437
	The Two-Factor Design	437
	The Concept of Interaction	441
	The Hypotheses That are Tested by a Two-Way Analysis of Variance	446
	Assumptions of the Two-Way Analysis of Variance	446
	Balanced versus Unbalanced Factorial Designs	448
	Partitioning the Total Sum of Squares	448
	Using the <i>F</i> -Ratio to Test the Effects in Two-Way ANOVA	449
	Carrying Out the Two-Way ANOVA Computation by Hand	450
	Decomposing Score Deviations about the Grand Mean	455
	Modeling Each Score As a Sum of Component Parts	455
	Explaining the Interaction As a Joint (or Multiplicative) Effect	456
	Measuring Effect Size	457
	Fixed versus Random Factors	459
	Post-Hoc Multiple Comparison Tests	459
	Simple Effects and Pairwise Comparisons	462
	Summary of Steps to Be Taken in a Two-Way ANOVA Procedure	465
	Conducting Power Analyses for Two-Way ANOVA	469
	Summary of R Commands	470
	Exercises	472
15	CORRELATION AND SIMPLE REGRESSION AS INFERENTIAL TECHNIQUES	481
	The Bivariate Normal Distribution	481
	Testing whether the Population Pearson Product-Moment Correlation Equals Zero	484
	Revisiting Simple Linear Regression for Prediction	488
	Estimating the Population Standard Error of Prediction, $\sigma_{Y X}$	489
	Testing the <i>b</i> -Weight for Statistical Significance	490
	Explaining Simple Regression Using an Analysis of Variance Framework	493
	Measuring the Fit of the Overall Regression Equation: Using <i>R</i> and R^2	495
	Relating R^2 to $\sigma^2_{Y X}$	496
	Testing R^2 for Statistical Significance	497
	Estimating the True Population R^2 : The Adjusted R^2	498
	Exploring the Goodness of Fit of the Regression Equation:	
	Using Regression Diagnostics	499
	Residual Plots: Evaluating the Assumptions Underlying Regression	500
	Detecting Influential Observations: Discrepancy and Leverage	507
	Using R to Obtain Leverage	509
	Using R to Obtain Discrepancy	509
	Using R to Obtain Influence	510
	Using Diagnostics to Evaluate the Ice Cream Sales Example	511
	Using the Prediction Model to Predict Ice Cream Sales	514
	Simple Regression When the Predictor Is Dichotomous	515
	Conducting Power Analyses for Correlation and Simple Regression	517
	Summary of R Commands	519
	Exercises	520

CONTENTS

xiii

16	AN INTRODUCTION TO MULTIPLE REGRESSION	531
	The Basic Equation with Two Predictors	532
	Equations for b , β , and $R_{Y.12}$ When the Predictors Are Not Correlated	533
	Equations for b , β , and $R_{Y.12}$ When the Predictors Are Correlated	534
	Summarizing and Expanding on Some Important Principles of Multiple Regression	536
	Testing The b -Weights for Statistical Significance	542
	Assessing the Relative Importance of the Predictor Variables in the Equation	543
	Measuring the Drop in R^2 Directly: An Alternative to the Squared Semipartial Correlation	544
	Evaluating the Statistical Significance of the Change in R^2	545
	The b -Weight As a Partial Slope in Multiple Regression	546
	Multiple Regression When One of the Two Predictor Variables Is Dichotomous	548
	Controlling Variables Statistically: A Closer Look	553
	A Hypothetical Example	554
	Conducting Power Analyses for Multiple Regression	556
	Summary of R Commands	557
	Exercises	559
17	TWO-WAY INTERACTIONS IN MULTIPLE REGRESSION	566
	Testing the Statistical Significance of an Interaction Using R	569
	Comparing the \hat{Y} Values from the Additive and Interaction Models	573
	Centering First-Order Effects if the Equation Has an Interaction	574
	Probing the Nature of a Two-Way Interaction	575
	Interaction When One of the Predictor Variables Is Dichotomous and the Other Is Continuous	578
	Methods Useful for Model Selection	583
	Summary of R Commands	586
	Exercises	588
18	NONPARAMETRIC METHODS	593
	Parametric versus Nonparametric Methods	593
	Nonparametric Methods When the Outcome Variable Is at the Nominal Level	594
	The Chi-Square Distribution (χ^2)	594
	The Chi-Square Goodness-of-Fit Test	597
	The Chi-Square Test of Independence	601
	Assumptions of the Chi-Square Test of Independence	604
	Fisher's Exact Test	606
	Calculating the Fisher's Exact Test by Hand Using the Hypergeometric Distribution	608
	Nonparametric Methods When the Outcome Variable Is Ordinal-Leveled	610
	Wilcoxon Sign Test	611
	The Mann-Whitney U -Test or Wilcoxon Rank-Sum Test	614
	The Kruskal-Wallis Analysis of Variance	617
	Summary of R Commands	619
	Exercises	620

<i>Appendix A</i>	<i>Dataset Descriptions</i>	626
<i>Appendix B</i>	<i>.R Files and Datasets in R Format</i>	641
<i>Appendix C</i>	<i>Statistical Tables</i>	642
<i>References</i>		664
<i>Index</i>		668

Preface

This first edition of *Statistics Using R: An Integrative Approach* capitalizes on the versatility and power of the R 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. R is a free, open-source software, which provides a command line approach to the analysis of data. This textbook teaches readers the skills necessary to program statistical analyses using this command line approach.

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 their 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, we have made available the datasets used in this textbook. The textbook's website contains the datasets used throughout the text. Also, we have published a software package, hosted on R's server, that can be downloaded as another means to access the datasets. This package, named *sur*, also contains several R functions that we have written to facilitate the application of several statistical methods.

The datasets used in this textbook provide examples of real-world data. Two are large sets of real data that we make repeated use of in both worked-out examples and end-of-chapter exercises. One dataset 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 datasets (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 datasets 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 functionality of R version 3.5.3 into the presentation of statistical material using a highly integrative approach that reflects practice. Students learn R 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 R within each chapter, all chapters have an associated .R file, that may be accessed from the textbook's website (www.cambridge.org/stats-r) along with other ancillary materials and resources. The .R 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 .R files by extracting and modifying commands from them. Emphasizing data workflow management throughout the text using the .R 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 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 nonparametric 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 18 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.

Finally, to facilitate learning R, an online tutorial has been written to accompany the textbook. The tutorial covers the basic use and manipulation of datasets, also referred to as data frames in R. The activities covered in the tutorial are designed to help students understand the examples in each chapter and to complete end-of-chapter exercises in the textbook. It also is designed to teach some basic coding skills that will be helpful when working with data frames in R, complete more complex analyses, and, ultimately, to learn the larger statistical concepts covered throughout the textbook.

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, Andrea Hassler, Jennifer Hill, 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.