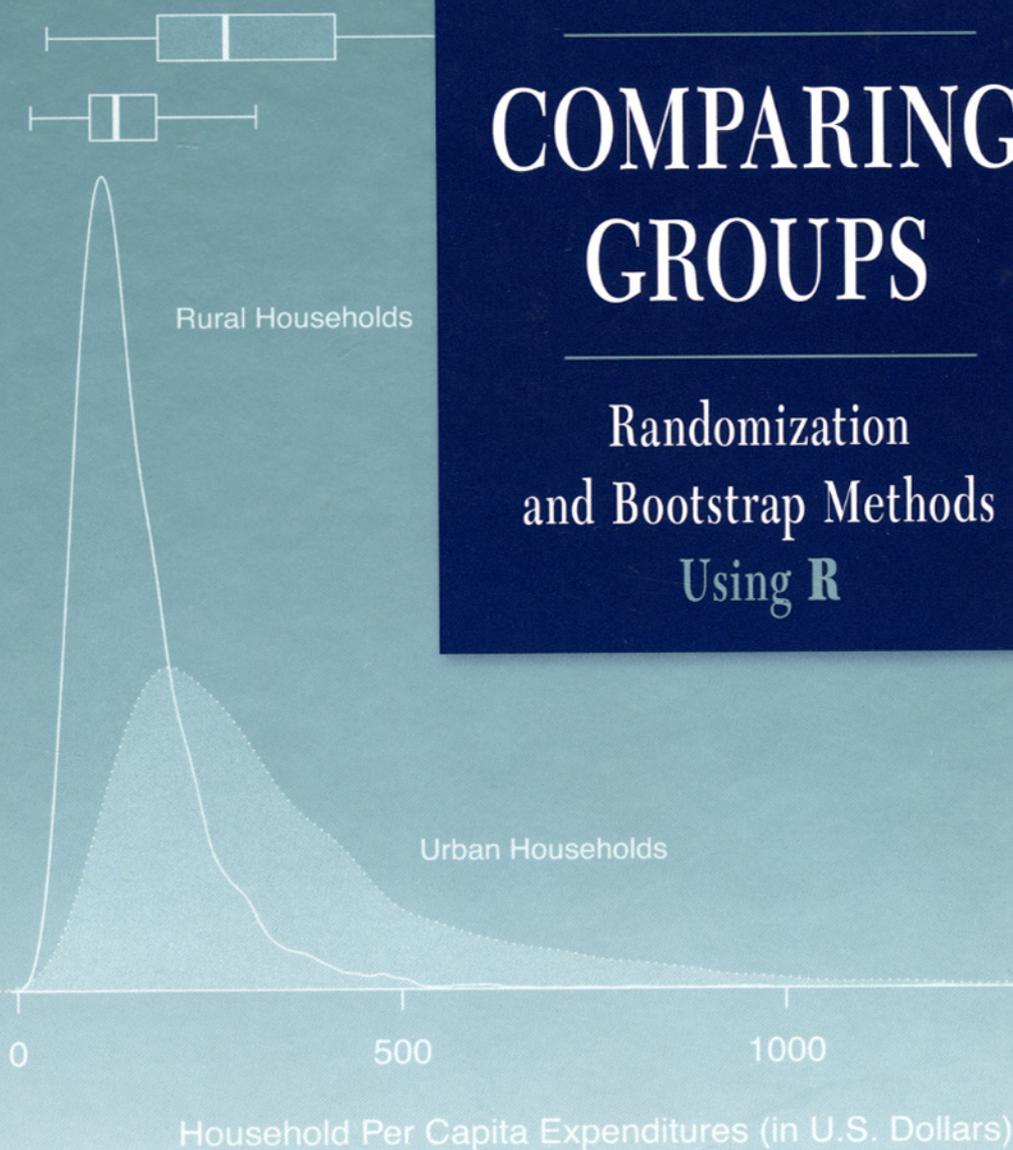


COMPARING GROUPS

Randomization
and Bootstrap Methods
Using **R**



Andrew S. Zieffler • Jeffrey R. Harring • Jeffrey D. Long

 **WILEY**

This page intentionally left blank

Comparing Groups

This page intentionally left blank

Comparing Groups

Randomization and Bootstrap Methods Using R

Andrew S. Zieffler

University of Minnesota

Jeffrey R. Harring

University of Maryland

Jeffrey D. Long

University of Iowa



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2011 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Zieffler, Andrew, 1974–

Comparing groups : randomization and bootstrap methods using R / Andrew Zieffler, Jeffrey Haring, Jeffrey D. Long.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-470-62169-1 (hardback)

1. Bootstrap (Statistics) 2. Random data (Statistics) 3. Psychology—Data processing. 4. R (Computer program language) 5. Distribution (Probability theory) I. Haring, Jeffrey, 1964– II. Long, Jeffrey D., 1964– III. Title.

QA276.8.Z53 2011

519.5'4—dc22

2010054064

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

List of Figures	xiii
List of Tables	xxi
Foreword	xxiii
Preface	xxv
Acknowledgments	xxx
1 An Introduction to R	1
1.1 Getting Started	2
1.1.1 Windows OS	2
1.1.2 Mac OS	2
1.1.3 Add-On Packages	2
1.2 Arithmetic: R as a Calculator	4
1.3 Computations in R: Functions	4
1.4 Connecting Computations	7
1.4.1 Naming Conventions	8
1.5 Data Structures: Vectors	9
1.5.1 Creating Vectors in R	9
1.5.2 Computation with Vectors	11
1.5.3 Character and Logical Vectors	12
	v

1.6	Getting Help	13
1.7	Alternative Ways to Run R	14
1.8	Extension: Matrices and Matrix Operations	14
1.8.1	Computation with Matrices	15
1.9	Further Reading	18
	Problems	19
2	Data Representation and Preparation	21
2.1	Tabular Data	23
2.1.1	External Formats for Storing Tabular Data	23
2.2	Data Entry	24
2.2.1	Data Codebooks	25
2.3	Reading Delimited Data into R	25
2.3.1	Identifying the Location of a File	26
2.3.2	Examining the Data in a Text Editor	28
2.3.3	Reading Delimited Separated Data: An Example	28
2.4	Data Structure: Data Frames	29
2.4.1	Examining the Data Read into R	29
2.5	Recording Syntax using Script Files	33
2.5.1	Documentation File	34
2.6	Simple Graphing in R	34
2.6.1	Saving Graphics to Insert into a Word-Processing File	35
2.7	Extension: Logical Expressions and Graphs for Categorical Variables	37
2.7.1	Logical Operators	38
2.7.2	Measurement Level and Analysis	40
2.7.3	Categorical Data	42
2.7.4	Plotting Categorical Data	44
2.8	Further Reading	45
	Problems	46
3	Data Exploration: One Variable	49
3.1	Reading In the Data	50
3.2	Nonparametric Density Estimation	52
3.2.1	Graphically Summarizing the Distribution	52
3.2.2	Histograms	52
3.2.3	Kernel Density Estimators	53
3.2.4	Controlling the Density Estimation	53

3.2.5	Plotting the Estimated Density	55
3.3	Summarizing the Findings	58
3.3.1	Creating a Plot for Publication	59
3.3.2	Writing Up the Results for Publication	61
3.4	Extension: Variability Bands for Kernel Densities	62
3.5	Further Reading	62
	Problems	63
4	Exploration of Multivariate Data: Comparing Two Groups	65
4.1	Graphically Summarizing the Marginal Distribution	66
4.2	Graphically Summarizing Conditional Distributions	66
4.2.1	Indexing: Accessing Individuals or Subsets	68
4.2.2	Indexing Using a Logical Expression	69
4.2.3	Density Plots of the Conditional Distributions	70
4.2.4	Side-by-Side Box-and-Whiskers Plots	70
4.3	Numerical Summaries of Data: Estimates of the Population Parameters	72
4.3.1	Measuring Central Tendency	73
4.3.2	Measuring Variation	74
4.3.3	Measuring Skewness	76
4.3.4	Kurtosis	78
4.4	Summarizing the Findings	80
4.4.1	Creating a Plot for Publication	80
4.4.2	Using Color	81
4.4.3	Selecting a Color Palette	85
4.5	Extension: Robust Estimation	87
4.5.1	Robust Estimate of Location: The Trimmed Mean	87
4.5.2	Robust Estimate of Variation: The Winsorized Variance	89
4.6	Further Reading	91
	Problems	91
5	Exploration of Multivariate Data: Comparing Many Groups	93
5.1	Graphing Many Conditional Distributions	94
5.1.1	Panel Plots	96
5.1.2	Side-by-Side Box-and-Whiskers Plots	97
5.2	Numerically Summarizing the Data	100
5.3	Summarizing the Findings	101
5.3.1	Writing Up the Results for Publication	102

5.3.2	Enhancing a Plot with a Line	102
5.4	Examining Distributions Conditional on Multiple Variables	103
5.5	Extension: Conditioning on Continuous Variables	107
5.5.1	Scatterplots of the Conditional Distributions	110
5.6	Further Reading	112
	Problems	113
6	Randomization and Permutation Tests	115
6.1	Randomized Experimental Research	118
6.2	Introduction to the Randomization Test	119
6.3	Randomization Tests with Large Samples: Monte Carlo Simulation	122
6.3.1	Rerandomization of the Data	124
6.3.2	Repeating the Randomization Process	125
6.3.3	Generalizing Processes: Functions	126
6.3.4	Repeated Operations on Matrix Rows or Columns	127
6.3.5	Examining the Monte Carlo Distribution and Obtaining the p -Value	127
6.4	Validity of the Inferences and Conclusions Drawn from a Randomization Test	130
6.4.1	Exchangeability	130
6.4.2	Nonexperimental Research: Permutation Tests	131
6.4.3	Nonexperimental, Nongeneralizable Research	131
6.5	Generalization from the Randomization Results	132
6.6	Summarizing the Results for Publication	133
6.7	Extension: Tests of the Variance	133
6.8	Further Reading	134
	Problems	135
7	Bootstrap Tests	137
7.1	Educational Achievement of Latino Immigrants	138
7.2	Probability Models: An Interlude	140
7.3	Theoretical Probability Models in R	141
7.4	Parametric Bootstrap Tests	143
7.4.1	Choosing a Probability Model	144
7.4.2	Standardizing the Distribution of Achievement Scores	144
7.5	The Parametric Bootstrap	146

7.5.1	The Parametric Bootstrap: Approximating the Distribution of the Mean Difference	146
7.6	Implementing the Parametric Bootstrap in R	148
7.6.1	Writing a Function to Randomly Generate Data for the boot() Function	148
7.6.2	Writing a Function to Compute a Test Statistic Using the Randomly Generated Data	150
7.6.3	The Bootstrap Distribution of the Mean Difference	151
7.7	Summarizing the Results of the Parametric Bootstrap Test	154
7.8	Nonparametric Bootstrap Tests	154
7.8.1	Using the Nonparametric Bootstrap to Approximate the Distribution of the Mean Difference	157
7.8.2	Implementing the Nonparametric Bootstrap in R	158
7.9	Summarizing the Results for the Nonparametric Bootstrap Test	160
7.10	Bootstrapping Using a Pivot Statistic	161
7.10.1	Student's t -Statistic	161
7.11	Independence Assumption for the Bootstrap Methods	164
7.12	Extension: Testing Functions	166
7.12.1	Ordering a Data Frame	166
7.13	Further Reading	168
	Problems	168
8	Philosophical Considerations	171
8.1	The Randomization Test vs. the Bootstrap Test	172
8.2	Philosophical Frameworks of Classical Inference	173
8.2.1	Fisher's Significance Testing	174
8.2.2	Neyman–Pearson Hypothesis Testing	175
8.2.3	p -Values	176
9	Bootstrap Intervals and Effect Sizes	179
9.1	Educational Achievement Among Latino Immigrants: Example Revisited	180
9.2	Plausible Models to Reproduce the Observed Result	180
9.2.1	Computing the Likelihood of Reproducing the Observed Result	181
9.3	Bootstrapping Using an Alternative Model	185
9.3.1	Using R to Bootstrap under the Alternative Model	187

9.3.2	Using the Bootstrap Distribution to Compute the Interval Limits	190
9.3.3	Historical Interlude: Student’s Approximation for the Interval Estimate	190
9.3.4	Studentized Bootstrap Interval	191
9.4	Interpretation of the Interval Estimate	191
9.5	Adjusted Bootstrap Intervals	192
9.6	Standardized Effect Size: Quantifying the Group Differences in a Common Metric	192
9.6.1	Effect Size as Distance—Cohen’s δ	193
9.6.2	Robust Distance Measure of Effect	195
9.7	Summarizing the Results	197
9.8	Extension: Bootstrapping the Confidence Envelope for a Q-Q Plot	197
9.9	Confidence Envelopes	198
9.10	Further Reading	202
	Problems	204
10	Dependent Samples	205
10.1	Matching: Reducing the Likelihood of Nonequivalent Groups	206
10.2	Mathematics Achievement Study Design	206
10.2.1	Exploratory Analysis	209
10.3	Randomization/Permutation Test for Dependent Samples	211
10.3.1	Reshaping the Data	212
10.3.2	Randomization Test Using the Reshaped Data	214
10.4	Effect Size	216
10.5	Summarizing the Results of a Dependent Samples Test for Publication	217
10.6	To Match or Not to Match . . . That is the Question	218
10.7	Extension: Block Bootstrap	220
10.8	Further Reading	223
	Problems	224
11	Planned Contrasts	227
11.1	Planned Comparisons	228
11.2	Examination of Weight Loss Conditioned on Diet	228
11.2.1	Exploration of Research Question 1	229
11.2.2	Exploration of Research Question 2	230

11.2.3	Exploration of Research Question 3	231
11.3	From Research Questions to Hypotheses	232
11.4	Statistical Contrasts	233
11.4.1	Complex Contrasts	236
11.5	Computing the Estimated Contrasts Using the Observed Data	237
11.6	Testing Contrasts: Randomization Test	239
11.7	Strength of Association: A Measure of Effect	240
11.7.1	Total Sum of Squares	241
11.8	Contrast Sum of Squares	243
11.9	Eta-Squared for Contrasts	243
11.10	Bootstrap Interval for Eta-Squared	244
11.11	Summarizing the Results of a Planned Contrast Test Analysis	245
11.12	Extension: Orthogonal Contrasts	245
11.13	Further Reading	251
	Problems	251
12	Unplanned Contrasts	253
12.1	Unplanned Comparisons	254
12.2	Examination of Weight Loss Conditioned on Diet	254
12.3	Omnibus Test	257
12.3.1	Statistical Models	257
12.3.2	Postulating a Statistical Model to Fit the Data	258
12.3.3	Fitting a Statistical Model to the Data	260
12.3.4	Partitioning Variation in the Observed Scores	262
12.3.5	Randomization Test for the Omnibus Hypothesis	268
12.4	Group Comparisons After the Omnibus Test	269
12.5	Ensemble-Adjusted p -values	270
12.5.1	False Discovery Rate	272
12.6	Strengths and Limitations of the Four Approaches	273
12.6.1	Planned Comparisons	273
12.6.2	Omnibus Test Followed by Unadjusted Group Comparisons	274
12.6.3	Omnibus Test Followed by Adjusted Group Comparisons	274
12.6.4	Adjusted Group Comparisons without the Omnibus Test	275
12.6.5	Final Thoughts	276
12.7	Summarizing the Results of Unplanned Contrast Tests for Publication	276

12.8	Extension: Plots of the Unplanned Contrasts	276
	12.8.1 Simultaneous Intervals	280
12.9	Further Reading	282
	Problems	283
References		285

LIST OF FIGURES

1.1	Example of the R console window on a Mac.	3
2.1	MRI scan stored as an image file.	22
2.2	Latino education data opened in <i>Text Editor</i> on the Mac. The data are delimited by commas and each case is delimited by a line break	24
2.3	Folder and file hierarchy on a computer system.	27
2.4	Box-and-whiskers plot of the distribution of educational achievement scores for 150 Latino immigrants.	36
2.5	Density plot of the distribution of educational achievement scores for 150 Latino immigrants.	36
2.6	Density plot of the distribution of educational achievement scores for 41 Latinos who immigrated to the United States between the ages of 4 and 7.	41
2.7	Bar plot showing the frequencies for Latinos in the sample who are fluent and not fluent in English.	45
2.8	Bar plot showing the proportions for Latinos in the sample who are fluent and not fluent in English.	45

3.1	Illustration of the kernel density estimation (solid line) for $N = 6$ observations (vertical lines). A Gaussian kernel function (dashed lines) with a fixed smoothing parameter was centered at each observation. The figure was adapted from Sain (1994).	54
3.2	Kernel density estimates using three different smoothing parameters. Each estimate used a Gaussian kernel. The estimate on the left used a smoothing parameter of 0.5. The estimate in the middle used a smoothing parameter of 2. The estimate on the right used a smoothing parameter of 10.	55
3.3	Kernel density estimates using Silverman’s rule of thumb for the smoothing parameter (left-hand plot). This is the default value for the smoothing parameter. The middle plot uses the Sheather and Jones “solve the equation” method of computing the smoothing parameter, and the right-hand plot uses an adaptive kernel function.	57
3.4	Kernel density estimate for the distribution of ages for a sample of 28,633 Vietnamese citizens using an adaptive kernel method.	59
3.5	Kernel density estimate (solid line) with variability bands (dotted lines) for the distribution of ages for the Vietnamese population. A smoothing parameter of $h = 4$ was used in the density estimation.	63
4.1	Kernel density estimate for the marginal distribution of household per capita expenditures (in U.S. dollars).	68
4.2	Kernel density plots for the distribution of household per capita expenditures (in dollars) conditioned on area. The density of household per capita expenditures for urban households is plotted using a solid line and that for rural households is plotted using a dotted line.	71
4.3	Side-by-side box-and-whiskers plots for the conditional distributions of household per capita expenditures (in U.S. dollars) using the <code>names=</code> argument.	72
4.4	Simulated density plots for the distribution of household per capita expenditures (in dollars) conditioned on area showing large within-group variation.	75
4.5	Simulated density plots for the distribution of household per capita expenditures (in dollars) conditioned on area showing small within-group variation.	75
4.6	Kernel density estimate for a mesokurtic distribution (solid line) and a leptokurtic distribution (dotted line). The leptokurtic distributions are skinnier and more peaked than the mesokurtic distribution.	79

4.7	Kernel density estimate for a mesokurtic distribution (solid line) and a platykurtic distribution (dotted line). The platykurtic distribution is flatter than the mesokurtic distribution.	79
4.8	Plots of the kernel density estimates for the distribution of household per capita expenditures (in dollars) for the rural ($n = 4269$) and urban households ($n = 1730$).	82
4.9	Kernel density plots for the distribution of household per capita expenditures (in dollars) conditioned on area. The density of household per capita expenditures for both distributions are shaded using the <code>polygon()</code> function. The plot appears in grayscale in this book, but should print in color on a computer.	84
5.1	Kernel density plots for the distribution of household per capita expenditures (in dollars) conditioned on region. The graph is problematic as the various regions are difficult to sort out. The plot appears in grayscale in this book, but should print in color on a computer.	96
5.2	The default panel array is <code>mfrow=c(1,1)</code> . This array has one row and one column.	97
5.3	This panel array is <code>mfrow=c(2,3)</code> . This array has two rows and three columns.	97
5.4	Panel plot of the density of household per capita expenditures (in dollars) conditioned on region.	99
5.5	Side-by-side box-and-whiskers plots of household per capita expenditures (in U.S. dollars) conditioned on region.	100
5.6	Plot of the estimated density of household per capita expenditures (in U.S. dollars) conditioned on region with the poverty line (shown as a dashed line) demarcated.	104
5.7	Plotted density estimates of the per capita household expenditures for the urban households (solid line) and rural households (dotted line) conditioned on region. The poverty line of \$119 is expressed as a vertical dashed line in each panel.	106
5.8	The distribution of mathematics achievement conditioned on average weekly time spent on mathematics homework. Achievement is reported as a T-score which are transformed scores having a mean of 50 and a standard deviation of 10.	110
5.9	Scatterplot showing the distribution of mathematics achievement conditioned on average weekly time spent on mathematics homework.	112

5.10	Scatterplot and side-by-side box-and-whiskers plots showing the distribution of mathematics achievement conditioned on average weekly time spent on mathematics homework.	112
6.1	Kernel density plots for the distribution of the T-scaled delinquency measure for students who participated in the after-school program (solid line) and for students who did not (dashed line).	124
6.2	Kernel density plot for the distribution of permuted mean differences. The point represents the observed mean difference of 1.7. The shaded area represents the p -value.	129
7.1	Density plots for the distribution of educational achievement conditioned on whether or not the Latin American country of emmigration was Mexico.	140
7.2	Side-by-side box-and-whiskers plots for the distribution of educational achievement conditioned on whether or not the Latin American country of emmigration was Mexico.	140
7.3	Spike plot of the probability distribution of the outcome for the number of heads in 10 tossed coins.	143
7.4	Density plots for the distribution of standardized educational achievement conditioned on whether or not the Latin American country of emmigration was Mexico.	147
7.5	Side-by-side box-and-whiskers plots for the distribution of standardized educational achievement conditioned on whether or not the Latin American country of emmigration was Mexico.	147
7.6	Steps for bootstrapping the distribution for a particular test statistic, \hat{V} , from a known probability distribution.	148
7.7	Visualization of the process used to parametrically bootstrap 150 observations from the standard normal distribution. The first 34 observations, the X replicates, constitute the replicated “non-Mexican immigrants” and the last 116 observations, the Y replicates, constitute the replicated “Mexican immigrants.” \hat{V} is the test statistic (e.g., the mean difference) computed for each replicate data set.	149
7.8	Bootstrap distribution of the mean difference in achievement scores between non-Mexican and Mexican immigrants for 4999 replicates under the hypothesis of no difference. The vertical line is drawn at 0.	153

- 7.9 Density plot (dashed line) with variability bands of the distribution of standardized educational achievement for Mexican immigrants. The solid line is the density curve of the standard normal probability model. 155
- 7.10 Density plot (dashed line) with variability bands of the distribution of standardized educational achievement for non-Mexican immigrants. The solid line is the density curve of the standard normal probability model. 155
- 7.11 Nonparametric bootstrapping involves resampling from the pooled observed sample with replacement. The bootstrap distribution of a statistic obtained from this method is typically a good approximation of the exact distribution. 156
- 7.12 Steps for using the sample data to bootstrap the distribution for a particular test statistic, \hat{V} , when the probability distribution is unknown. 157
- 7.13 Bootstrap distribution of the mean difference (using nonparametric bootstrapping) in achievement scores between non-Mexican and Mexican immigrants for 4999 replicates assuming there are no population differences. The vertical line is drawn at zero. 159
- 9.1 The bootstrap distributions of the standardized mean difference based on the model with parameter value 0 (dotted), parameter value 3 (dashed), and parameter value 10 (dot-dashed). The point in the plot demarcates the observed mean difference in the data of 5.9. 182
- 9.2 The bootstrap distribution of the mean difference based on the model with parameter value 3. A vertical dashed line is drawn at 3, the model's parameter value. The point in the plot demarcates the observed mean difference in the data of 5.9. The one-sided p -value, indicating the likelihood of this model to reproduce the observed data, is shaded. 184
- 9.3 Nonparametric bootstrapping under an alternative model involves resampling from the observed sample with replacement within defined strata. 187
- 9.4 Steps for using the sample data to bootstrap the distribution for a particular test statistic, \hat{V} , under an alternative model. 188
- 9.5 Bootstrap distribution of the mean difference in educational achievement scores between non-Mexican and Mexican immigrants for 4999 replicates. The resampling was carried out under the alternative model, thus centering the distribution around the observed mean difference of 5.9 (dashed line). 190

9.6	Density plot of the 4999 bootstrapped robust standardized effect sizes. The resampling was carried out under the alternative model, thus centering the distribution around the observed robust standardized effect of 0.38 (vertical dashed line).	198
9.7	Q-Q plot of the standardized Mexican achievement scores. The line references where the points would lie if the sample of scores were normally distributed.	199
9.8	Density plot of the empirical standardized Mexican achievement scores. The normal probability model is shown with the dashed line.	199
9.9	Q-Q plot of the bootstrapped scores drawn in the first replicate. The line references where the points would lie if the sample of scores was consistent with a normal distribution.	201
9.10	Q-Q plot of the bootstrapped standardized scores for $R = 25$ replicates of $n = 116$ resampled observations from the standard normal distribution. The bold line references where the points would lie if the sample of scores was consistent with a normal distribution.	202
9.11	Q-Q plot of the studentized Mexican achievement scores along with a 95% confidence envelope. The confidence envelope was computed using $R = 4999$ replicates of $n = 116$ observations resampled using a parametric bootstrap from the $N(0, 1)$ distribution.	203
9.12	Density plot of the empirical standardized Mexican achievement scores. The confidence envelope based on the normal probability model is also added to the plot.	203
10.1	Distribution of PSAT scores conditioned on treatment.	209
10.2	Density plots for the distribution of mathematics achievement scores conditioned on treatment.	211
10.3	Side-by-side box-and-whiskers plots for the distribution of mathematics achievement scores conditioned on treatment.	211
10.4	Randomization distribution of the mean difference under the null hypothesis of no difference. A vertical line (solid) is drawn at 0, the hypothesized mean difference.	216
10.5	Randomization distribution of the mean difference for both the assumptions of independence and dependence. A vertical line (dotted) is drawn at the hypothesized mean difference of 0.	219

10.6	The steps used to block bootstrap the distribution for a particular test statistic, \hat{V} , under the alternative model.	220
10.7	Bootstrap distribution of the standardized mean difference as measured by Glass's delta in mathematics achievement scores between treatment and control participants for 4999 replicates. The resampling was carried out using block bootstrapping under the alternative model. The vertical line (dotted) is drawn at the observed standardized mean difference of 0.51.	223
11.1	Conditional density plots showing the distribution of weight loss for the Atkins and the Ornish diets.	231
11.2	Conditional density plots showing the distribution of weight loss for the Atkins and the other three diets.	233
11.3	Conditional density plots showing the distribution of weight loss for the Atkins/Zone and the LEARN/Ornish diets.	235
11.4	Steps used to carry out a randomization test to obtain the distribution for a particular contrast under the null model of no difference.	239
11.5	Side-by-side box-and-whiskers plots showing the distribution of word recall conditioned on time spent studying. The conditional means are each marked with a triangle. The dotted line, which connects the conditional means, shows a potential polynomial trend.	249
12.1	Side-by-side box-and-whiskers plots of 12-month weight change conditioned on diet.	256
12.2	The deviation between the observed score and the marginal mean (horizontal line) for four observations.	263
12.3	The deviation between the observed score and the marginal mean (horizontal line) for four observations. Each has now been decomposed into the part of the deviation that can be explained by diet (solid) and that which cannot (dashed).	264
12.4	Graphical representation of the squared deviations from the marginal mean weight change for four observations.	265
12.5	Graphical representation of the squared decomposed model deviations for four observations—the deviation between the conditional mean (dotted horizontal line) and the marginal mean (solid horizontal line). The squares represent the squared model deviations for the four observations.	266

12.6	Graphical representation of the squared decomposed residual deviations for four observations—the deviation between the conditional mean (dotted horizontal line) and the observed value. The squares represent the squared residual deviations for the four observations.	267
12.7	Steps to carry out a randomization test to obtain the distribution for the omnibus effect under the null hypothesis.	268
12.8	Dynamite plot showing the conditional mean weight change and standard error for each diet.	278
12.9	Plot showing the point estimate and bootstrap interval estimate of the mean difference for each contrast based on the bias-corrected-and-accelerated adjustment using 4999 bootstrap replicates.	280
12.10	Plot showing the point estimate and bootstrap interval estimate of the mean difference for each contrast based on the Benjamani–Hochberg adjusted significance level. Each interval uses the bias-corrected-and-accelerated adjustment using 4999 bootstrap replicates. A vertical line is plotted at zero, the parameter value associated with no difference.	283

LIST OF TABLES

2.1	Comparison Operators Used in R	38
2.2	Logical Operators Used in R	39
4.1	Common Colors and Their Associated RGB Values Using a Maximum Color Value of 255	83
4.2	Small Data Set (original data) Showing Results of Trimming (trimmed data) and Winsorizing (winsorized data) Distribution by 20%	88
4.3	Comparison of Conventional and Robust Estimates of Location and Variation for Conditional Distributions of Household per Capita Expenditures	91
5.1	Mean, Standard Deviation and Sample Sizes for Household per Capita Expenditures (in dollars) Conditioned on Region ^a	103
5.2	Mean, Standard Deviation and Sample Sizes for Household per Capita Expenditures (in dollars) Conditioned on Urbanicity and Region	105
6.1	Four Potential Scenarios Researcher Could Face When Making Inferences	117
		xxi

6.2	Ten Possible Permutations of the Perceived Social Competence Scores and Difference in Mean Perceived Social Competence Scores	121
7.1	Probability Distribution of Random Variable X for Number of Heads When Tossing 10 Coins	142
7.2	Results of Carrying Out Nonparametric Bootstrap Test for Three Different Test Statistics ^a	164
8.1	Four Potential Scenarios Researcher Could Face When Making Inferences ^a	173
8.2	Fisher’s Descriptions for Scale of Evidence Against the Null Hypothesis ^a	177
9.1	Parameter Value, p -Value, and Qualitative Description of Strength of Evidence Against the Model for 11 Potential Models	185
10.1	Data for Participants in First Two Blocks Presented in Long Format	212
10.2	Data for Participants in First Two Blocks Presented in Wide Format	213
12.1	Mean and Standard Deviation for 12-month Weight Change (in kg) Conditioned on Diet. ^a	256
12.2	Contrast Coefficients, Observed Contrast Value, Randomization Test Results, Eta-Squared Estimate, and Bootstrap Results for Interval Estimate of Eta-Squared for Each of Six Pairwise Contrasts ^a	271
12.3	Randomization Test Results for Both Unadjusted Method and Benjamani–Hochberg Ensemble-Adjusted Method ^a	276
12.4	Computation of Adjusted Significance Level for Benjamani–Hochberg Simultaneous Confidence Intervals Based on Sequential Unadjusted p -Values	281
12.5	Unadjusted and Simultaneous Bootstrap Intervals for Each of Six Contrasts ^a	282

FOREWORD

This book is truly different.

It is rare that a textbook with so few technical prerequisites offers so much of value, not just to its intended student audience, but also to all practicing statisticians. The book in your hands is one of those rare ones. It is the product—over years—of careful study and deep thought by its authors, who are practicing statisticians, education researchers, and root-deep innovators.

Statistics, its teaching and its learning, have seen a number of exciting developments over the last quarter century. Among these:

1. Near-universal acceptance of the assertion that applied statistics textbooks should emphasize the priority of real data over mathematically driven abstract exposition.
2. A coming-of-age for computer-intensive methods of data analysis.
3. A parallel recognition by teachers of statistics that, along with an emphasis on real data and reliance on applied context, computer simulation offers an alternative to abstract derivations as an approach to exposition and understanding.
4. A cooperative development and broad embrace of R, the public domain, and open-source software, whose growing community of contributors has made cutting edge, computationally intensive methods freely available around the world.
5. A coming-of-age of statistics education as a research area in its own right, and, as a result, a growing clarity about what is hardest for students to learn, and a sharpened focus on what is most important for students to learn.

6. A change at the foundations of our subject: What we can compute shapes what we can do; what we can do shapes what we actually do in practice; what we do in practice shapes how we think about what we do. (As just one example, when I first was learning statistics in the 1960s, Bayesian analyses were typically dismissed as mere idealism; today, the combination of hierarchical models and Markov chain Monte Carlo has turned Bayes resistors into old fogeys.)

By integrating these developments in a novel way, *Comparing Groups* deserves to be regarded as one of this quarter-century's pioneering textbooks aimed at introducing nonstatisticians to contemporary thought and practice.

Many features are distinctive:

1. The exposition is visual/intuitive/computational rather than haunted by the old headless horseman of abstract derivations.
2. The writing is strikingly good—the exposition reflects the authors' careful attention to word choice. (In my experience, it is rare in statistical exposition to read with a sense that the writers made thoughtful, deliberate choices about wording after considering a variety of options.)
3. The references are a gold mine.
4. The use of modern graphics is unusual for a book at this level. I refer, in particular, to panel plots and kernel density estimates.
5. Prerequisites are minimal. No calculus is needed. Although it helps to know about vectors and matrices, one does not need a linear algebra course.
6. The emphasis is practical throughout. For example, the exposition refers systematically to APA guidelines and offers sample write-ups.
7. Content reflects the current research literature. For example, the exposition recognizes the importance of effect sizes, and the treatment of multiple testing addresses the recent research on false discovery rates.
8. Overall, the emphasis is on statistics with a purpose, statistics for deep interpretive understanding.

In short, this is a book that reduces prerequisites, avoids technical baggage, focuses on essentials, and teaches via authentic applications. It offers readers our profession's current best sense of how to understand data for comparing groups, taking full advantage of computationally intensive methods such as randomization and the bootstrap.

George W. Cobb

Robert L. Rooke Professor of Statistics, emeritus, Mount Holyoke College

Vernon Wilson Endowed Visiting Professor, Eastern Kentucky University

PREFACE

Computational advances have changed the face of statistical practice by transforming what we do and by challenging how we think about scientific problems.

—R. A. Thisted & P. F. Velleman (1992)

Drawing conclusions and inferences about the differences among groups is an almost daily occurrence in the scholarly life of an educational or behavioral researcher. Group comparisons are at the heart of many research questions addressed by these researchers. These are questions about efficacy, such as “Is a particular curriculum effective in improving students’ achievement?” and questions about magnitude such as “How much lower are attendance rates for a particular population of students?”

The content in this book provides the statistical foundation for researchers interested in answering these types of questions through the introduction and application of current statistical methods made possible through computation—including the use of Monte Carlo simulation, bootstrapping, and randomization tests. Rather than focus on mathematical calculations like so many other introductory texts in the behavioral sciences, the approach taken here is to focus on conceptual explanations and the use of statistical computing. We agree with the sentiments of Moore (1990, p. 100), who stated, “calculating sums of squares by hand does not increase understanding; it merely numbs the mind.”

At the heart of every chapter there is an emphasis on the direct link between research questions and data analysis. Purposeful attention is paid to the integration of design, statistical methodology, and computation to propose answers to research questions based on appropriate analysis and interpretation of quantitative data. Practical suggestions for analysis and the presentation of results based on suggestions from the *APA Publication Manual* are also included. These suggestions are intended to help researchers clearly communicate the results of a data analysis to their audience.

Computation as a Tool for Research

Computation is ubiquitous in everyday life. This is in large part due to the progress made in technology in the last 20 years. The price of computational power continually becomes less expensive and more powerful. In a recent *New York Times* article Caleb Chung—the inventor of the electronic Furby toy—was quoted as saying, “the price of processing power has dropped to the floor. I can buy the equivalent of an Apple II processor for a dime” (Marriot, 2007, p. 9).

Computing has become an essential part of the day-to-day practice of statistical work. It has not only greatly changed the practice of statistics itself, but has influenced the development of new state-of-the-art statistical methodologies and broadened the types of questions that can now be addressed by research scientists applying these newly derived data analytic techniques. Previous to having the availability of such computational power, the outcome of a research project may have been crucially dependent on the efficiency or numerical accuracy of algorithms employed by a methodologist. Computational advantages have allowed educational and behavioral researchers to capitalize on sheer computing power to solve problems that previously were impracticable or intractable.

Although computer-based data analysis probably covers most of the activity that educational and behavioral researchers use in their work, statistical computing is more than just using a software package for the analysis of data. It also encompasses the programming and development of new functionality or software, the analysis of statistical and numerical algorithms, and data management. Furthermore, computing is an important avenue for allowing statisticians and scientists to pursue lines of inquiry that in the past would not have been possible. It is also an essential tool in any modern collaborative effort.

To support and help facilitate the use of scientific computing, examples using the R computer language will be used throughout the book. Rather than relegate examples to the end of chapters, our approach is to interweave the computer examples with the narrative of the monograph. R is free, which means it is available to individuals with various resources. It is also a professional-level software environment that implements many of the modern statistical methods emphasized in the monograph. Furthermore, the architecture of the object orientation in R easily allows data analysis to be performed based on structured inquiry, in which a series of interrelated and increasingly complex questions are posed to guide and inform the process of data analysis. In the classroom, structured inquiry can increase the opportunities for

intellectual engagement (e.g., Badley, 2002; Cooper, 2005; Lee, 2005; Prince & Felder, 2006). Lastly, since R is a functional programming language, it can be easily extended for use in a great many analytic scenarios in the behavioral and educational sciences.

Organization of the Book

The organization of this book has been shaped by the authors' experiences in teaching the statistical and computing content presented in many graduate courses for social science students. The topics introduced represent the authors' beliefs about relevant content for inclusion in an introductory graduate-level statistics course for educational and behavioral science students.

This content covers statistical computing, exploratory data analysis, and statistical inference. The individual chapters in the book endeavor to integrate these ideas to help guide educational and behavioral researchers through the process of data analysis when making group comparisons. While most chapters address multiple strands of the data analytic process, in surveying the computing and statistical content of each chapter, it is clear that each chapter is primarily focused to address a particular strand.

Statistical Computing

The first two chapters are written primarily for educational and behavioral researchers who may not have prior experience with computing environments and, in particular, with the R statistical computing environment. Chapter 1 introduces the fundamentals of R. Basic ideas of computation are presented as the building blocks for the remainder of the monograph. In Chapter 2, these ideas are further developed through the reading in and manipulation of external data (data saved outside of R). This chapter also introduces script files as a way of recording useful syntax.

Exploratory Data Analysis

Chapters 3, 4, and 5 introduce graphical and numerical methods of exploratory analysis for group comparisons. These chapters begin the integration of the statistical and computing content. Chapter 3 focuses on graphical exploration of a marginal distribution to introduce the statistical and computing content without the additional cognitive load of making comparisons. Ideas of kernel density estimation are at the forefront of this chapter. Additionally, the use of R to create publication quality plots is highlighted.

Chapters 4 and 5 focus on the exploration of conditional distributions for two and more than two groups, respectively. Graphical exploration is expanded to facilitate comparisons (e.g., panel plots) and numerical summary measures are introduced to quantify characteristics of the distributions.

Statistical Inference

Chapter 6 is the first of two chapters on the use of Monte Carlo methods for statistical inference. This chapter presents randomization and permutation tests to examine group differences. Chapter 7, the second such chapter, presents parametric and nonparametric bootstrap tests. Both of these chapters highlight the quantification of how likely an observed statistic is given the expected variation in that statistic because of chance. Chapter 6 highlights the expected random variation due to random assignment, whereas Chapter 7 highlights the random variation due to random sampling.

Chapter 8 expands on the differences between the randomization and bootstrap tests by offering philosophical reasons for using each method apart from the design employed by the researcher. This is especially highlighted for observational studies. This chapter also offers an overview of differences between the Fisher and Neyman–Pearson philosophies of statistical testing.

Chapters 9 and 10 expand the realm of statistical inference. Chapter 9 introduces ideas of interval estimation. The bootstrap methodology is developed further in this chapter, but under stratification of groups. Standardized and unstandardized measures of effect size are discussed. In Chapter 10, testing and estimation for dependent samples is introduced. Rather than do this through the use of repeated measures, which is often the case in statistics books in the educational and behavioral sciences, dependence is introduced via the use of blocking in the research design.

Chapters 11 and 12 introduce common methods in the educational and behavioral sciences for comparing more than two groups. Chapter 11 is focused on planned comparisons. Exploration, testing and estimation are revisited in the context of multiple groups. The use of linear contrasts is introduced to facilitate these ideas.

Chapter 12 is focused on unplanned comparisons. Three common methods are presented: (1) adjusted group comparisons without the omnibus test; (2) unadjusted group comparisons following an omnibus test; and (3) adjusted group comparisons following an omnibus test. Strengths, weaknesses, and criticisms of each method are discussed. The use of ensemble-adjusted p -values and adjusted bootstrap intervals are discussed in this chapter.

Extras

This book refers to and uses several data sets throughout the text. Each of these data sets is available online at <http://www.tc.umn.edu/~zief0002/ComparingGroups.html>. The codebook for each data set is also available at the Web site. Many of these data sets were used in actual research studies that have been published and were graciously provided by the authors and researchers involved in those studies.

The R script files for each chapter are also available at the URL mentioned above. The commands are provided so that the reader can reproduce all the output and plots discussed in the monograph.

Problems can be found at the end of each chapter, except Chapter 8. These problems are meant to provide a platform to perform data analyses using real data