



# Data Analysis by Resampling: Concepts and Applications

Clifford E. Lunneborg

D U X B U R Y P R E S S



0212-2  
L963

# Data Analysis by Resampling

---

## Concepts and Applications

**Clifford E. Lunneborg**

University of Washington



**Duxbury**  
Thomson Learning™

---

Australia • Canada • Mexico • Singapore • Spain • United Kingdom • United States

Sponsoring Editor: Carolyn Crockett  
Marketing Team: Tom Ziolkowski, Beth Kroenke,  
Laura Hubrich  
Editorial Assistant: Ann Day  
Production Coordinator: Kelsey McGee

Production Service: Robin Gold/Forbes Mill Press  
Permissions Editor: Mary Kay Hancharick  
Cover Design: Christine Garrigan  
Cover Printing, Printing, and Binding: R. R. Donnelley  
and Sons/Crawfordsville

COPYRIGHT © 2000 by Brooks/Cole  
Duxbury is an imprint of Brooks/Cole, a division of Thomson Learning  
The Thomson Learning logo is a trademark used herein under license.

For more information, contact:  
DUXBURY  
511 Forest Lodge Road  
Pacific Grove, CA 93950 USA  
[www.duxbury.com](http://www.duxbury.com)

All rights reserved. No part of this work may be reproduced, transcribed or used in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, Web distribution, or information storage and/or retrieval systems—without the prior written permission of the publisher.

For permission to use material from this work, contact us by  
Web: [www.thomsonrights.com](http://www.thomsonrights.com)  
fax: 1-800-730-2215  
phone: 1-800-730-2214

Printed in United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Lunneborg, Clifford E.

Data analysis by resampling : concepts and applications / Clifford

E. Lunneborg

p. cm.

Includes bibliographic references and index.

ISBN 0 534-22110-6 (case)

1. Resampling (Statistics) I. Title.

QA278.8.L86 2000

99-40284

519.5--dc21



# Data Analysis by Resampling

---

Concepts and Applications



E200000806



[www.duxbury.com](http://www.duxbury.com)

Valuable resources @ no additional charge

## Preface

---

# Data Analysis by Resampling

This book discusses, from the perspective of modern statistical computing, those questions empirical researchers most commonly ask of their case-based studies. Increases in computer power and decreases in computer costs have revolutionized the role of computing in statistics. One important aspect to this is the growing importance of data analyses based on recycling the scores constituting a data set, a collection of techniques broadly known as resampling. These computer-intensive techniques repeat a data analysis many times on replicate data sets (the resamples), all based on an observed set of data.

Why are resampling techniques important? Until inexpensive computing power made replicate data analysis practical, the drawing of statistical inferences from a set of data almost always required that we accept an idealized model for the origin of those data. Such models can be either inappropriate or inadequate for the data in our study. Resampling techniques allow us to base the analysis of a study solely on the design of that study, rather than on a poorly-fitting model.

How have models dictated data analysis? Here is a prime example. To compare the responses of cases (e.g., patients, students, or animals) receiving different treatments (e.g., alternative anti-inflammatory drugs, reading instruction, or fat-controlled diets), researchers commonly have adopted an analysis of variance (ANOVA) model for the data. Briefly, ANOVA models the responses to treatment as those of cases that were: (a) randomly sampled from a population or populations in which the distribution of responses to each treatment would (b) have the same or homogeneous variance and (c) follow that of a normal random variable

The ANOVA model leads to simple data analysis computations.

More often than not, at least one of the assumptions of the ANOVA model will be palpably or, at least, arguably untrue. To take an example, we assess the reading abilities of students participating in our study comparing methods of reading instruction with a test yielding integer scores between 0 and 50. How might the ANOVA model be inappropriate to these data?

1. The population distribution of such scores could not be normal, that is, it could not include all magnitudes between negative and positive infinity.
2. Although the population distributions might have a common variance, we could never know this from our sampled data. All we can expect to do is to identify some data sets for which an underlying population homogeneity of variance is highly unlikely.
3. Finally, of course, our cases might not have been sampled randomly from any student population, rather they might have been all the first-graders enrolled last autumn in Shady Brook primary school.

The resampling inference techniques featured in this book either eliminate or restrict to a minimum any unwarranted or unverifiable assumptions about our data. These techniques depend almost exclusively on the data and on our knowledge of how the data were collected—our knowledge of the design of the study. Our resampling data analyses, then, can be described as design-based rather than model-based.

Our resampling techniques let us perform data analyses when we believe one or more of the assumptions of a classical, model-based technique are inappropriate.

The range of classical statistical models is limited. The ANOVA model, to pursue this important example, allows the researcher to estimate confidence intervals for contrasts among population distribution means. A mean contrast, though, might not be the researcher's choice for a treatment comparison. A researcher might prefer a confidence interval for a population effect size measure, such as  $d$  which scales the difference in means for a control and treatment population by the standard deviation of the control population. But there is no classical model for estimating a confidence interval for this effect size measure. Our resampling methods adapt to the questions that researchers regard as important, rather than requiring that researchers ask only questions for which there are model-based answers.

We characterized the resampling approaches to statistical inference as design-based. In this book, we develop and illustrate three distinct resampling approaches, targeting the three major and distinct data collection designs employed in case-based studies:

1. Bootstrap Population Inference for studies in which cases (patients, animals, customers, student subjects, classrooms, clinics, agricultural or forest tracts, and so on) have been randomly sampled from one or more populations.

2. Rerandomization Causal Inference for studies in which cases, though not randomly chosen, have been randomly allocated among two or more treatments.
3. Subsample Descriptive Inference for studies in which cases have been neither randomly sampled nor randomly allocated.

This book has two audiences. The first comprises graduate and advanced undergraduate students interested in the design and analysis of empirical research, perhaps their own. Students should have had an introduction to statistical ideas at the level of standard errors, confidence intervals, and tests of statistical hypotheses.

The second audience consists of active researchers, trained in classical, model-based, statistical methodology but who are now interested in learning how to use the more realistic data analyses that modern statistical computing has made available.

## **Goals**

The major goals of the book are the following:

1. Establish the importance of matching the data analysis for a study to
  - a. the sampling design of the study
  - b. the amount of data collected
  - c. the precision of the data collected
  - d. what is known about any populations sampled
  - e. what the researcher wants to learn from the study
2. Introduce three resampling approaches to data analysis that satisfy these important criteria.
3. Show how each of these resampling approaches can be applied across the range of questions commonly asked of case-based empirical data.
4. Provide explicit algorithms for a wide variety of bootstrap, rerandomization, and subsampling applications.
5. Illustrate the use in resampling analyses of three widely available statistical packages.

A critical and unique feature of this book is the firm linkage developed between study design and analysis. A number of otherwise very good resampling books have appeared in recent years (e.g., Davison & Hinkley,

1997, Efron & Tibshirani, 1993, Good, 1994, Manly, 1997, and Sprent, 1998) that unfortunately blur the distinctions that should be drawn among the principal case-selection designs—random sampling, randomization, and nonrandom.

In addition, the range of applications treated here, extending from one- and two-group designs through more complex multiple treatment designs to linear and logistic regression models, is well suited to the needs of researchers conducting case-based studies.

## **Organization**

The book is organized into two parts.

### ***Part I – Resampling Concepts with Applications***

Concepts 1 through 16 begin with a review of basic ideas in statistical inference and then introduce the three approaches to resampling inference. These Concepts are short, each the basis for a single lecture, and are accompanied by Applications that provide examples and exercises to illustrate the Concepts.

### ***Part II – Resampling Applications by Study Design***

Applications 17 through 23 develop resampling applications for the more common case-based study designs. These applications begin with the single data set, move through analyses of multiple independent data sets or treatments and repeated measures, and finish with techniques for linear models with measured or categorical response data. At each stage, we outline the limitations of classical inference and develop applications of each of the three resampling approaches.

A final Postscript summarizes resampling approaches to statistical inference.

## ***Selection of Topics***

Instructors might want to choose among the Concepts and Applications, or to give greater emphasis to some than to others, depending on the amount of instructional time available, the background of students, and the goals of the course.

- Students having completed a previous course at the level of Moore and McCabe's *Introduction to the Practice of Statistics* will be able to review quickly the material in Concepts and Applications 1 through 5 and have



time in a one-semester course to survey, at least, the full range of Applications.

- For students less-well grounded in classical statistical inference, instructors can give more time to the initial Concepts and, where time is short, limit the course to introducing resampling ideas, extending the course perhaps no further than Concepts and Applications 16.
- Where more time is available, some instructors might choose to emphasize Applications 19 through 21, if the goal is to develop a nonparametric alternative to the analysis of variance, whereas others might feature Applications 22 and 23, if linear models are the central concern.

In my experience, resampling course instruction is divided profitably and fairly equally between two activities, one-half of each week's meetings given to lecture and discussion of reading, the other half to directed (exemplar) and free (assigned homework) computer lab work. It is important to provide, early on, ample time for students to acquaint themselves with any software that is new to them.

The bootstrap, rerandomization, and subsampling all use random sequences of resamples. As a result, students working on the same problem will obtain slightly different answers. A computational exercise that I have found valuable early in the course demonstrates how the variability in student answers decreases as the number of resamples is increased. Students can learn in this way how to choose an appropriate number of resamples.

## **Computational Support**

Resampling is computer-intensive. Applications of the techniques presented in this book require some programming in a statistical computing package. Some statistical packages make resampling easier than others. This text uses three statistical computing packages that have features well adapted to resampling computations, S-Plus (MathSoft, 1997), SC (Dusoir, 1997) and Resampling Stats (Bruce, Simon, & Oswald, 1995). These are not the only possibilities, and students, their instructors, or their institutions might have other well-established statistical computing preferences.

Where possible, adopting one of the three featured packages to support a course taught from this text seems natural. But another package certainly can be used. In Part I, logical, step-by-step computational algorithms are presented for each resampling application in turn. These are then illustrated in

enough variety in Part II that students should be able to program any application in another package with which they are proficient.

The text illustrates the use of three statistical packages. We don't have enough room, however, to provide a primer or introduction to the use of those packages. If students are not already familiar with a package, instructors should provide some additional support in getting students started. Each of the three featured packages has excellent documentation, and S-Plus has a number of short third party getting-started texts.

## Accessibility

The mathematical presentation is at the level of college algebra. This reflects the fact that the resampling approaches to data analysis do not depend on deep mathematical reasoning. In particular, no knowledge of calculus is required. There are a few brief algebraic developments, associated with the definition and estimation of confidence intervals and with hypothesis testing. In Applications 22, the  $Q$ -statistic, a nonparametric version of the F-test, is presented using the notation of linear algebra. Because the necessary computational details are easily encapsulated, readers unfamiliar with matrix operations should not be disadvantaged in applying the approach. The linear log-odds models described in Applications 23 will be understood more easily if the student is familiar with logarithmic and antilogarithmic transformations. All the student needs by way of mathematical preparation, beyond college algebra, is an understanding of probability as taught in the typical introductory applied statistics course. The functional notation for cumulative distribution functions (cumulative proportions or probabilities) used to unify the resampling approaches might be new to some readers and therefore is introduced with some care.

## Examples

The text includes a fairly large number of real-data examples and exercises. These were chosen to illustrate: (a) the three basic case-selection types—random samples, random allocation, and nonrandom; (b) a variety of experimental designs; and (c) a range of substantive content.

One invaluable source of data has been the publication, *A Handbook of Small Data Sets* (Hand, D. J. et al., 1994). Instructors searching for additional examples might find what they are looking for in that excellent collection.

## Acknowledgments

My understanding of and approach to resampling and statistical inference have been shaped over the years by contact and correspondence with a number of statisticians actively working in the area. These include Gordon Bear, Peter Bruce, Steve Buckland, Brian Cade, David Draper, Tony Dusoier, Eugene Edgington, Bradley Efron, David Hand, Andrew Hayes, Tim Hesterberg, Chris Jones, Cyrus Mehta, Patrick Onghena, Normand Péladeau, Galen Shorack, Bruce Thompson, Robert Tibshirani, Jon Wellner, and Bruno Zumbo. They are not responsible, singly or collectively, for the choices I made and the resulting form of the book.

More recently, I am greatly indebted to the reviewers of earlier drafts of this book: Peter Bruce of Resampling Stats, Inc., Christine Anderson-Cook of Virginia Polytechnic Institute and State University, Bruce Thompson of Texas A&M University and Baylor College of Medicine, Rodney Wolff of Queensland University of Technology, and Arthur Yeh of Bowling Green State University. Their thoughtful critiques have been vital to shaping the final content and form of this book. Finally, three cheers for my editors at Duxbury-Brooks/Cole—Alex Kugushev, Carolyn Crockett, and Kelsey McGee—and for Robin Gold at Forbes Mill Press for turning draft into reality.

## Table of Contents

### Part I: Resampling Concepts

<b>Introduction</b>		<b>1</b>
<b>Concepts 1:</b>	<b>Terms and Notation</b>	<b>5</b>
	Cases, Attributes, Scores, and Treatments	5
	Experimental and Observational Studies	7
	Data Sets, Samples, and Populations	7
	Parameters, Statistics, and Distributions	8
	Distribution Functions	9
<b>Applications 1:</b>	<b>Cases, Attributes, and Distributions</b>	<b>13</b>
	Attributes, Scores, Groups, and Treatments	13
	Distributions of Scores and Statistics	13
	Exercises	14
<b>Concepts 2:</b>	<b>Populations and Random Samples</b>	<b>19</b>
	Varieties of Populations	19
	Random Samples	24
<b>Applications 2:</b>	<b>Random Sampling</b>	<b>27</b>
	Simple Random Samples	27
	Exercises	30
<b>Concepts 3:</b>	<b>Statistics and Sampling Distributions</b>	<b>33</b>
	Statistics and Estimators	33
	Accuracy of Estimation	34
	The Sampling Distribution	34
	Bias of an Estimator	37
	Standard Error of a Statistic	37
	RMS Error of an Estimator	38
	Confidence Interval	38
<b>Applications 3:</b>	<b>Sampling Distribution Computations</b>	<b>43</b>
	Exercises	43
<b>Concepts 4:</b>	<b>Testing Population Hypotheses</b>	<b>45</b>
	Population Statistical Hypotheses	45

	Population Hypothesis Testing	46
<b>Applications 4:</b>	<b>Null Sampling Distribution p-Values</b>	<b>51</b>
	The p-Value of a Directional Test	52
	The p-Value of a Nondirectional Test	53
	Exercises	55
<b>Concepts 5:</b>	<b>Parametrics, Pivotal, and Asymptotics</b>	<b>57</b>
	The Unrealizable Sampling Distribution	57
	Sampling Distribution of a Sample Mean	58
	Parametric Population Distributions	60
	Pivotal Form Statistics	60
	Asymptotic Sampling Distributions	63
	Limitations of the Mathematical Approach	65
<b>Applications 5:</b>	<b>CI for Normal Population <math>\mu</math> and <math>\sigma^2</math></b>	<b>67</b>
	CI for a Normal Population Mean	67
	CI for a Normal Population Variance	69
	Nonparametric CI Estimation	71
	Exercises	72
<b>Concepts 6:</b>	<b>Limitations of Parametric Inference</b>	<b>75</b>
	Range and Precision of Scores	75
	Size of Population	76
	Size of Sample	76
	Roughness of Population Distributions	77
	Parameters and Statistics of Interest	77
	Scarcity of Random Samples	78
	Resampling Inference	78
<b>Applications 6:</b>	<b>Resampling Approaches to Inference</b>	<b>83</b>
	Exercises	83
<b>Concepts 7:</b>	<b>The Real and Bootstrap Worlds</b>	<b>85</b>
	The Real World of Population Inference	85
	The Bootstrap World of Population Inference	87
	Real-World Population Distribution Estimates	88
	Nonparametric Population Estimates	96
	Sample Size and Distribution Estimates	98
<b>Applications 7:</b>	<b>Bootstrap Population Distributions</b>	<b>99</b>
	Nonparametric Population Estimates	99
	Exercises	103
<b>Concepts 8:</b>	<b>The Bootstrap Sampling Distribution</b>	<b>105</b>
	The Bootstrap Conjecture	105

	Complete Bootstrap Sampling Distributions	106
	Monte Carlo Bootstrap Distributions	108
	The Bootstrap Estimate of Standard Error	110
	The Bootstrap Estimate of Bias	111
	Simple Bootstrap CI Estimates	112
<b>Applications 8:</b>	<b>Bootstrap SE, Bias, and CI Estimates</b>	<b>119</b>
	Example	119
	Exercises	122
<b>Concepts 9:</b>	<b>Better Bootstrap CIs: The Bootstrap-<math>t</math></b>	<b>125</b>
	Pivotal Form Statistics	125
	The Bootstrap- $t$ Pivotal Transformation	127
	Forming Bootstrap- $t$ CIs	129
	Estimating the Standard Error of an Estimate	132
	Range of Application of the Bootstrap- $t$	135
	Iterated Bootstrap CIs	136
<b>Applications 9:</b>	<b>SEs and CIs for Trimmed Means</b>	<b>137</b>
	Definition of the Trimmed Mean	137
	Importance of the Trimmed Mean	138
	A Note on Outliers	139
	Determining the Trimming Fraction	140
	Sampling Distribution of the Trimmed Mean	141
	Applications	143
	Exercise	154
<b>Concepts 10:</b>	<b>Better Bootstrap CIs: BCA Intervals</b>	<b>157</b>
	Bias-Corrected and Accelerated CI Estimates	157
	Applications of the BCA CI	166
	Better Confidence Interval Estimates	166
<b>Applications 10:</b>	<b>Using CI Correction Factors</b>	<b>167</b>
	Requirements for a BCA CI	167
	Implementations of the BCA Algorithm	168
	Exercise	179
<b>Concepts 11:</b>	<b>Bootstrap Hypothesis Testing</b>	<b>181</b>
	CIs, Null Hypothesis Tests, and p-Values	181
	Bootstrap- $t$ Hypothesis Testing	186
	Bootstrap Hypothesis Testing Alternatives	187
	CI Hypothesis Testing	190
	Confidence Intervals or p-Values?	193
<b>Applications 11:</b>	<b>Bootstrap p-Values</b>	<b>195</b>
	Computing a Bootstrap- $t$ p-Value	195

	Fixed- $\alpha$ CIs and Hypothesis Testing	198
	Computing a BCA CI p-Value	200
	Exercise	202
<b>Concepts 12:</b>	<b>Randomized Treatment Assignment</b>	<b>205</b>
	Two Functions of Randomization	205
	Randomization of Sampled Cases	206
	Randomization of Available Cases	206
	Statistical Basis for Local Causal Inference	207
	Population Hypotheses Revisited	213
<b>Applications 12:</b>	<b>Monte Carlo Reference Distributions</b>	<b>217</b>
	Serum Albumen in Diabetic Mice	217
	Resampling Stats Analysis	219
	SC Analysis	222
	S-Plus Analysis	223
	Exercises	224
<b>Concepts 13:</b>	<b>Strategies for Randomizing Cases</b>	<b>227</b>
	Independent Randomization of Cases	227
	Completely Randomized Designs	231
	Randomized Blocks Designs	233
	Restricted Randomizations	235
	Constraints on Rerandomization	236
<b>Applications 13:</b>	<b>Implementing Case Rerandomization</b>	<b>237</b>
	Completely Randomized Designs	237
	Randomized Blocks Designs	243
	Independent Randomization of Cases	248
	Restricted Randomization	251
	Exercises	254
<b>Concepts 14:</b>	<b>Random Treatment Sequences</b>	<b>257</b>
	Between- and Within-Cases Designs	257
	Randomizing the Sequence of Treatments	258
	Causal Inference for Within-Cases Designs	260
	Sequence Randomization Strategies	265
<b>Applications 14:</b>	<b>Rerandomizing Treatment Sequences</b>	<b>269</b>
	Analysis of the AB-BA Design	269
	Sequences of $K > 2$ Treatments	276
	Exercises	285
<b>Concepts 15:</b>	<b>Between- and Within-Cases Designs</b>	<b>289</b>
	Between/Within Designs	289
	Between/Within Resampling Strategies	293

	Doubly Randomized Available Cases	294
<b>Applications 15:</b>	<b>Interactions and Simple Effects</b>	<b>299</b>
	Simple and Main Effects	299
	Exercise	308
<b>Concepts 16:</b>	<b>Subsamples: Stability of Description</b>	<b>309</b>
	Nonrandom Studies and Data Sets	309
	Local Descriptive Inference	310
	Description Stability and Case Homogeneity	311
	Subsample Descriptions	312
	Employing Subsample Descriptions	316
	Subsamples and Randomized Studies	318
<b>Applications 16:</b>	<b>Structured and Unstructured Data</b>	<b>319</b>
	Half-Samples of Unstructured Data Sets	319
	Subsamples of Source-Structured Cases	323
	Exercises	325

## Part II: Resampling Applications

<b>Introduction</b>		<b>331</b>
<b>Applications 17:</b>	<b>A Single Group of Cases</b>	<b>335</b>
	Random Sample or Set of Available Cases	335
	Typical Size of Score Distribution	334
	Variability of Attribute Scores	342
	Association Between Two Attributes	349
	Exercises	356
<b>Applications 18:</b>	<b>Two Independent Groups of Cases</b>	<b>359</b>
	Constitution of Independent Groups	359
	Location Comparisons for Samples	361
	Magnitude Differences, CR and RB Designs	368
	Magnitude Differences, Nonrandom Designs	372
	Study Size	375
	Exercises	381
<b>Applications 19:</b>	<b>Multiple Independent Groups</b>	<b>385</b>
	Multiple Group Parametric Comparisons	385
	Nonparametric $K$ -Group Comparisons	389
	Comparisons among Randomized Groups	393
	Comparisons among Nonrandom Groups	401
	Adjustment for Multiple Comparisons	404
	Exercises	407



<b>Applications 20:</b>	<b>Multiple Factors and Covariates</b>	<b>411</b>
	Two Treatment Factors	411
	Treatment and Blocking Factors	423
	Covariate Adjustment of Treatment Scores	424
	Exercises	440
<b>Applications 21:</b>	<b>Within-Cases Treatment Comparisons</b>	<b>447</b>
	Normal Models, Univariate and Multivariate	447
	Bootstrap Treatment Comparisons	449
	Randomized Sequence of Treatments	452
	Nonrandom Repeated Measures	455
	Exercises	458
<b>Applications 22:</b>	<b>Linear Models: Measured Response</b>	<b>465</b>
	The Parametric Linear Model	465
	Nonparametric Linear Models	474
	Prediction Accuracy	498
	Linear Models for Randomized Cases	501
	Linear Models for Nonrandom Studies	503
	Exercises	506
<b>Applications 23:</b>	<b>Categorical Response Attributes</b>	<b>513</b>
	Cross-Classification of Cases	513
	The $2 \times 2$ Table	529
	Logistic Regression	531
	Exercises	546
<b>Postscript:</b>	<b>Generality, Causality, and Stability</b>	<b>551</b>
	Study Design and Resampling	552
	Resampling Tools	556
<b>References</b>		<b>557</b>
<b>Index</b>		<b>561</b>