

A Biostatistics Toolbox for Data Analysis

This sophisticated package of statistical methods is for advanced master's degree and Ph.D. students in public health and epidemiology who are involved in the analysis of data. It makes the link from statistical theory to data analysis, focusing on the methods and data types most common in public health and related fields. Like most toolboxes, the statistical tools in this book are organized into sections with similar objectives. Unlike most toolboxes, however, these tools are accompanied by complete instructions, explanations, detailed examples, and advice on relevant issues and potential pitfalls – conveying skills, intuition, and experience.

The only prerequisite is a first-year statistics course and familiarity with a computing package such as R, Stata, SPSS, or SAS. Though the book is not tied to a particular computing language, its figures and analyses were all created using R. Relevant R code, data sets, and links to public data sets are available from www.cambridge.org/9781107113084.

STEVE SELVIN is a professor of biostatistics in the School of Public Health at the University of California, Berkeley, and was the head of the division from 1977 to 2004. He has published more than 250 papers and authored several textbooks in the fields of biostatistics and epidemiology. His book *Survival Analysis for Epidemiologic and Medical Research* was published by Cambridge University Press in 2008.

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Frontmatter
[More information](#)

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Frontmatter
[More information](#)

A Biostatistics Toolbox for Data Analysis

Steve Selvin
University of California, Berkeley



Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Frontmatter
[More information](#)

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

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/9781107113084

© Steve Selvin 2015

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 2015

Printed in the United States of America

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

Library of Congress Cataloging in Publication Data

Selvin, S.

A biostatistics toolbox for data analysis / Steve Selvin, University of California, Berkeley.

pages cm

Includes bibliographical references and index.

ISBN 978-1-107-11308-4 (hardback)

1. Medical statistics. 2. Biometry. I. Title.

RA409.S327 2015

610.2'1 – dc23 2015012679

ISBN 978-1-107-11308-4 Hardback

Additional resources for this publication at www.cambridge.org/9781107113084

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
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Frontmatter
[More information](#)

For Nancy, David, Liz, Ben, and Eli

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Frontmatter
[More information](#)

Contents

<i>Preface</i>	xv
Basics	
1 Statistical Distributions	3
The Normal Probability Distribution	3
The t -Distribution	7
The Chi-Square Probability Distribution	9
Illustration of the Pearson Chi-Square Statistic – Simplest Case	12
The f -Distribution	13
The Uniform Probability Distribution	13
The p -value	15
A Few Relationships among Probability Distributions	16
2 Confidence Intervals	19
Four Properties of an Estimated Confidence Interval	20
Confidence Intervals for a Function of an Estimate	21
Confidence Intervals Based on Estimates near Zero	21
Exact Confidence Interval (Computer-Estimated)	22
A Confidence Interval Based on an Estimated Median Value	23
A Confidence Interval and a Confidence Band	25
Details: A Confidence Interval and a Statistical Test	27
3 A Weighted Average	29
A Basic Application	29
Ratios and Weighted Averages	30
Estimates Weighted by Reciprocal Variances	32
A Puzzle	37
Age-Adjusted Rates Using Weighted Averages	39
Smoothing – A Weighted Average Approach	40
Example: Weighted Average Smoothing of Hodgkin’s Disease Mortality Data	44
4 Two Discrete Probability Distributions	47
Binomial Probability Distribution	49

	Two Applications of the Binomial Distribution	53
	The Geometric Probability Distribution	57
	A Poisson Probability Distribution	58
	Two Applications of the Poisson Probability Distribution	62
	A Note on Rare Events	67
5	Correlation	69
	Spearman's Rank Correlation Coefficient	73
	Point Biserial Correlation Coefficient	74
	Nonparametric Measure of Association: The γ -Coefficient	77
	A Special Case: The $2 \times k$ Table	80
	Chi-Square-Based Measures of Association	81
	Proportional Reduction in Error Criterion	83
	Applications	
6	The 2×2 Table	87
	The Analysis of a 2×2 Table	89
	Measures of Association in a 2×2 Table	92
	Odds Ratio and Relative Risk Ratio	94
	A Correction to Improve the Normal Distribution Approximation	96
	The Hypergeometric Probability Distribution	98
	Fisher's Exact Test	101
	Correlation in a 2×2 Table	102
	A 2×2 Table with a Structural Zero	106
	Assessing the Accuracy of a Diagnostic Test	107
7	Linear Bivariate Regression Model	111
	The Bivariate Linear Regression Model	111
	Additivity	114
	Coefficients	114
	Multiple Correlation Coefficient	116
	Adjustment	118
	Interaction	119
	Confounder Bias	121
	Collinearity	123
8	The $2 \times k$ Table	126
	Wilcoxon (Mann-Whitney) Rank Sum Test	126
	Nonparametric Analysis of a $2 \times k$ Table	132
	A Chi-Square Analysis of a $2 \times k$ Table	134
	Another Example: Childhood Cancer and X-Ray Exposure	137
9	The Loglinear Poisson Regression Model	141
	A Simple Poisson Regression Model	141

Contents

ix

	Poisson Regression Model: Analysis of Vital Statistics Data	144
	Rate Ratios Adjusted for Several Variables	146
	A Test for Linear Trend	148
	A Graphical Display of a Table	150
	Implementation of Poisson Models Applied to Tabular Data	151
	Analysis of Tables with Incomplete Data	153
	Another Illustration of the Analysis of an Incomplete Table	154
	Quasi-independence: Analysis of an Incomplete Table	156
	A Case Study – Adjusted Mortality Rates: Black-White Infant Mortality	158
	First Approach: Weight-Specific Comparisons	161
	Second Approach: A Model-Free Summary	163
	Third Approach: Poisson Regression Model	165
10	Two-Way and Three-Way Tables and Their Analysis	170
	Analysis of Tables – Continuous Data	177
	Matched Pairs – The Categorical Variable Case	182
	Analysis of Tables – Count Data	185
	Three-Way Tables	189
	The Analysis of the Three-Way Table	190
	Complete Independence	190
	Joint Independence	191
	Conditional Independence	191
	No Pairwise Independence	192
	Log-Linear Models for Four $2 \times 2 \times 2$ Three-Way Tables	193
	Example of Completely Independent Variables	193
	Example of Jointly Independent Variables	194
	Example of Conditional Independent Variables	195
	Example of Additive Relationships	196
	A Case Study – Joint Independence	197
	A Case Study – Conditional Independence in a $2 \times 2 \times 4$ Table	201
11	Bootstrap Analysis	204
	Example: Analysis of Paired Data	209
	Example: Evaluation of a Difference between Two Harmonic Mean Values	211
	An Example of Bootstrap Estimation from Categorical Data	214
	Kappa Statistic – A Bootstrap-Estimated Confidence Interval	216
	A Graphic Application of Bootstrap Estimation	218
	A Property of Bootstrap Estimation	220
	Randomization and Bootstrap Analysis Applied to Two-Sample Data	220
	Randomization Test	222
	Bootstrap Analysis	223

12	Graphical Analysis	227
	A Confidence Interval and a Boxplot	227
	Multiple Comparisons – A Visual Approach	229
	The Cumulative Probability Distribution Function	229
	Inverse Functions for Statistical Analysis	232
	Kolmogorov One-Sample Test	241
	Kolmogorov-Smirnov Two-Sample Test	243
	Simulation of Random “Data” with a Specific Probability Distribution	245
	Simulation of “Data” with a Continuous Probability Distribution	248
	Simulation of “Data” with a Discrete Probability Distribution	249
	A Case Study – Poisson Approximation	249
	A Graphical Approach to Regression Analysis Diagnostics	250
	A Graphical Goodness-of-Fit for the Logistic Regression Model	257
13	The Variance	260
	A Brief Note on Estimation of the Variance	260
	A Confidence Interval	261
	Test of Variance	262
	Homogeneity/Heterogeneity	264
	Analysis of Variance – Mean Values	265
	Another Partitioning of Variability	268
	Two-Sample Test of Variance	269
	<i>F</i> -Ratio Test of Variance	270
	Bartlett’s Test of Variance	271
	Levene’s Test of Variance	273
	Siegel-Tukey Two-Sample Test of Variance	274
	Comparison of Several Variances	275
14	The Log-Normal Distribution	278
	Example: Lognormal Distributed Data	281
	A Left-Censored Log-Normal Distribution	283
	An Applied Example	284
15	Nonparametric Analysis	287
	The Sign Test	289
	The Wilcoxon Signed Rank Test	291
	Kruskal-Wallis Nonparametric Comparison of <i>k</i> Sample Mean Values	294
	Three-Group Regression Analysis	297
	Tukey’s Quick Test	300
	Friedman Rank Test	302
	Survival	
16	Rates	309
	An Average Mortality Rate	309

Contents

xi

	An Approximate Average Rate	314
	A Rate Estimated from Continuous Survival Times	318
17	Nonparametric Survival Analysis	324
	Cumulative Hazard Function	333
	Description of the Median Survival Time	334
	Comparison of Two Survival Probability Distributions – The Log-Rank Test	335
	Proportional Hazards Rates	339
	An Example of Survival Analysis	342
	The Cox Analysis of Proportional Hazards Models	345
	Proportional Hazards Model – A Case Study	346
18	The Weibull Survival Function	352
	Two-Sample Comparison – The Weibull Model	358
	The Shape Parameter	360
	Multivariable Weibull Survival Time Model – A Case Study	363
	Epidemiology	
19	Prediction, a Natural Measure of Performance	371
	Net Reclassification Index (Binary Case)	373
	Net Reclassification Index (Extended Case)	375
	Integrated Discrimination Improvement	379
	Summary Lines as Measures of Improvement in Model Prediction	380
	Covariance and Correlation	382
	Epilogue	383
	Least Squares Estimation – Details	383
20	The Attributable Risk Summary	387
	Incidence Rates	391
	Two Risk Factors	393
	Adjustment by Stratification	394
	Adjustment by Model	395
	Issues of Interpretation	396
	A Few Less Specific Issues of Interpretation	397
21	Time-Space Analysis	399
	Knox’s Time-Space Method	399
	The Statistical Question Becomes: Is $m = 0$?	400
	Test of Variance Applied to Spatial Data	403
	Mantel’s Time-Space Regression Method	404
	Performance and Time-Space Methods	405

22	ROC Curve and Analysis	407
	An ROC Curve	407
	An ROC Analysis Example	414
	Nonparametric Estimation of an ROC Curve	417
	An Example – Nonparametric ROC Analysis	418
	How to Draw a Nonparametric ROC Curve	419
	Construction of the ROC Curve	419
	Implementation	425
	Genetics	
23	Selection: A Statistical Description	433
	Stable Equilibrium	434
	A Description of Recombination	435
	Two Examples of Statistical Analyses	439
	Selection – Recessive Lethal	442
	A More General Selection Pattern	444
	Selection – A Balanced Polymorphism	445
	Fitness	448
24	Mendelian Segregation Analysis	450
	Ascertainment	450
	Truncation	451
	Estimation of a Segregation Probability: Complete Ascertainment	451
	Estimation of a Segregation Probability: Single Ascertainment	455
25	Admixed Populations	458
	A Model Describing Admixture	459
	Estimation of the Admixture Rate	461
	An Example: Estimation of the Extent of Admixture	462
26	Nonrandom Mating	465
	Genotype Frequencies – Correlation and Variance	465
	Genetic Variance	469
	Wahlund's Model	470
	Random Genetic Drift	473
	A Description: Consecutive Random Sampling of a Genetic Population	473
	Random Genetic Drift – A Description	474
	The Statistics of Random Genetic Drift	474
	Mutation/Selection Equilibrium	477
	The Story of Mr. A and Mr. B	477
	One-Way Mutation	479
	Mutation/Selection Balance	480

	<i>Contents</i>	xiii
	Assortative Mating	481
	Assortative Mating Model	482
	Theory	
27	Statistical Estimation	489
	The Sample	489
	Covariance	490
	The Population	492
	Infinite Series with Statistical Applications	496
	Example of Infinite Power Series	497
	Binomial Theorem	498
	Functions of Estimates	499
	Approximate Values for the Expectation and Variance of a Function	500
	Partitioning the Total Sum of Squares	503
	Expected Value and Variance of Wilcoxon Signed Rank Test	504
	Maximum Likelihood Estimation	505
	Properties of a Maximum Likelihood Estimate	510
	Example Maximum Likelihood Estimates	512
	Method of Moments Estimation	515
	<i>Appendix: R Code</i>	519
	<i>Index</i>	557

Cambridge University Press
978-1-107-11308-4 - A Biostatistics Toolbox for Data Analysis
Steve Selvin
Frontmatter
[More information](#)

Preface

Books on statistical methods largely fall into two categories: elementary books that describe statistical techniques and mathematically advanced texts that describe the theory underlying these techniques. This text is about the art and science of applying statistical methods to the analysis of collected data and provides an answer to the question, after a first course in statistics, What next? Like most toolboxes, the statistical tools in the text are loosely organized into sections of similar objectives. Unlike most toolboxes, these tools are accompanied by complete instructions, explanations, detailed examples, and advice on relevant issues and potential pitfalls. Thus, the text is a sophisticated introduction to statistical analysis and a necessary text for master's and Ph.D. students in epidemiology programs as well as others whose research requires the analysis of data.

The text employs a pattern of describing sampled data and providing examples followed by a discussion of the appropriate analytic strategies. This approach introduces the reader to data and analytic issues and then explores the logic and statistical details. A large number of examples and illustrations are included to appeal to a diverse audience. It is often the case that examples produce substantial insight into the problem in hand. Most statistical texts reverse this approach and describe the statistical method first followed by examples.

The level of this text is beyond introductory but far short of advanced.

Fundamental topics, such as the median values, simple linear regression methods, correlation coefficients, and confidence intervals found in most introductory books are not repeated but frequently reinforced. The explanations, illustrations, and examples require little or no mathematics to completely understand the presented material. Nevertheless, simple mathematical notation and arguments are included because they are unambiguous, and frequently their clarity leads to better understanding.

The complexity of the mathematics that supports the discussion of the statistical tools is no more than high school algebra. Mostly the symbols and notation from mathematics are used to describe the various approaches to data analysis. An exception is the last chapter, which is intended to enrich the applied nature of the text with a bit of statistical theory.

The choices of the statistical tools included in the text are largely based on two criteria: the usefulness in the analysis of data from human populations as well as a variety of other methods because they simply demonstrate general statistical data analysis strategies. The level of presentation of the techniques discussed evolved from a second-year course in biostatistics for epidemiology graduate students taught over the last decade at the University of California, Berkeley. Also, much of the material has been presented in summer courses at the Graduate Summer Session in Epidemiology at the University of Michigan School of Public Health and more recently at The Johns Hopkins Bloomberg School of Public Health as

part of the Summer Institute of Epidemiology and Biostatistics. In other words, the material has been thoroughly “classroom tested.”

The text is organized in six sections.

Section I: BASICS is a review of fundamental statistical distributions and methods: for example, confidence intervals and correlation including important details not typically included in introductory texts.

Section II: APPLICATIONS builds on these basic statistical tools to create more advanced strategies used to analyze specific kinds of issues and data. Linear models, contingency table analysis, graphical methods, and bootstrap estimation are examples of methods that are frequently useful in unraveling the sometimes complicated issues that data are collected to explore. Also included are parallel and often neglected nonparametric methods.

The next three sections continue to develop new methods as well as include further extensions of basic statistical tools but applied to specific subject matter areas: survival, epidemiologic, and genetic data. These sections are a “proving grounds” for statistical techniques applied to a variety of challenging kinds of data.

Section III: SURVIVAL contains discussions of three important statistical techniques designed for the analysis of survival time data. Specifically, they consist of an extensive exploration of rates and their properties, followed by descriptions of nonparametric methods and regression models specifically designed to analyze survival data.

Section IV: EPIDEMIOLOGY similarly explores statistical methods particularly designed for analysis of epidemiologic data such as attributable risk analysis, cluster analysis, ROC curves, and reclassification methods.

Section V: GENETICS presents statistical tools applied to several fundamental topics to give a sense of the role of statistics and data analysis in a genetics context. Topics include a statistical description of selection/mutation dynamics, sibship analysis, and application to several ways that statistical tools identify the consequences of nonrandom mating.

Section VI: THEORY is focused on the underlying principles of a few statistical tools frequently used in data analysis settings. This last section, however, is not about data analysis but is a “user friendly” introduction to how important data analysis tools work. Many “tools” of modern society are effectively used without even a hint of how and why they work. Statistical methods are not an exception. This readily accessible “beginner’s guide to statistical theory” is intended to enrich the focus on applications. Primarily, these theoretical details demystify the sometimes baffling expressions that appear when the focus is entirely on description and application. With surprisingly little effort, such mysterious expressions and the sometimes obscure logic of statistical techniques disappear with use of high school algebra and a bit of first semester calculus.

Key Features

The text examples mostly consist of small subsets of real data so that the analytic results described are easily verified or explored with alternative analyses. These examples also range over a number of methods and kinds of data, from estimating missing values to special

techniques potentially useful in the analysis of genetic data. Several statistical methods are illustrated with the analysis of complete and again real data in terms of comprehensive case studies.

The introduction of many of the statistical methods discussed begins with an example made up of only a few artificial observations. These miniature examples allow simple “hands-on” illustrations of the more complicated computational and technical details. In addition, more than 150 figures provide visual displays of analytic concepts, and issues adding yet another important dimension to the use of statistical tools to analyze data. The distinguished statistician John W. Tukey stated, “The greatest value of a picture is it forces us to notice what we never see.”

It is not possible to characterize statistical methods in a definite linear or logical sequence. The discussed techniques are indexed to create a “road map” so that the interconnections among the discussed material can be traced back and forth throughout the text.

Nonparametric methods are typically treated as a separate topic. These methods are integrated into the text where they are natural partners to parallel parametric methods. These important techniques enrich the practice and understanding of most statistical approaches and are particularly important when small sample sizes are encountered.

The text does not contain references to parallel published material. The ability to use online searches to locate references certainly replaces the tradition of annotating a text. Using Google, for example, easily produces not only specific material referenced in the text but a wide range of other perhaps worthwhile sources on the same topic.

An appendix contains the computer code that produced the worked-out examples in the text. The statistical software used is entitled simply “R.” Free and easily downloaded, it provides an extensive statistical analysis system. The explanations and details in the appendix are not comprehensive and are left to manuals and books created exactly for this purpose. The R system, however, is completely self-documenting with direct access to descriptions of the R language, computer code, and examples. In addition, this popular data analysis system is well documented in numerous books, and a huge number of online computer sites exist that provide specific instructions and illustrations. The R analyses in the appendix can be used in variety of ways in conjunction with the text. First, using the presented code to verify the text material provides an extremely detailed description of the application of specific methods and statistical approaches. Furthermore, with minor modification, additional data sets can be analyzed to furnish alternative examples or assess results from larger sample sizes or gauge the influences of extreme values. The text does not refer to or depend on the R code, and the appendix can be completely ignored. At the other extreme, this appendix presents the opportunity to be part of learning a useful statistical analysis software language. R is available from www.r-project.org.

Remember, the study of statistics makes it possible to acquire the ability to state with great certainty the degree of uncertainty.