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.

A Biostatistics Toolbox for Data Analysis

Steve Selvin
University of California, Berkeley

CAMBRIDGE
UNIVERSITY PRESS

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

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

For Nancy, David, Liz, Ben, and Eli

Contents

Genetics

Preface

Books on statistical methods largely fall into two categories: elementary books that describe statistical techniques and mathematically advanced texts that describe the theory underling 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.