Biostatistics

A Methodology for the Health Sciences

Second Edition

GERALD VAN BELLE
LLOYD D. FISHER
PATRICK J. HEAGERTY
THOMAS LUMLEY

Department of Biostatistics and
Department of Environmental and
Occupational Health Sciences
University of Washington
Seattle, Washington

A JOHN WILEY & SONS, INC., PUBLICATION
Biostatistics
WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: David J. Balding, Noel A. C. Cressie, Nicholas I. Fisher,
Iain M. Johnstone, J. B. Kadane, Geert Molenberghs, Louise M. Ryan,
David W. Scott, Adrian F. M. Smith, Jozef L. Teugels
Editors Emeriti: Vic Barnett, J. Stuart Hunter, David G. Kendall

A complete list of the titles in this series appears at the end of this volume.
Ad majorem Dei gloriam
# Contents

Preface to the First Edition ix
Preface to the Second Edition xi

1. Introduction to Biostatistics 1
2. Biostatistical Design of Medical Studies 10
3. Descriptive Statistics 25
4. Statistical Inference: Populations and Samples 61
5. One- and Two-Sample Inference 117
6. Counting Data 151
7. Categorical Data: Contingency Tables 208
9. Association and Prediction: Linear Models with One Predictor Variable 291
10. Analysis of Variance 357
11. Association and Prediction: Multiple Regression Analysis and Linear Models with Multiple Predictor Variables 428
12. Multiple Comparisons 520
13. Discrimination and Classification 550
14. Principal Component Analysis and Factor Analysis 584
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.</td>
<td>Rates and Proportions</td>
<td>640</td>
</tr>
<tr>
<td>16.</td>
<td>Analysis of the Time to an Event: Survival Analysis</td>
<td>661</td>
</tr>
<tr>
<td>17.</td>
<td>Sample Sizes for Observational Studies</td>
<td>709</td>
</tr>
<tr>
<td>18.</td>
<td>Longitudinal Data Analysis</td>
<td>728</td>
</tr>
<tr>
<td>19.</td>
<td>Randomized Clinical Trials</td>
<td>766</td>
</tr>
<tr>
<td>20.</td>
<td>Personal Postscript</td>
<td>787</td>
</tr>
<tr>
<td></td>
<td>Appendix</td>
<td>817</td>
</tr>
<tr>
<td></td>
<td>Author Index</td>
<td>841</td>
</tr>
<tr>
<td></td>
<td>Subject Index</td>
<td>851</td>
</tr>
<tr>
<td></td>
<td>Symbol Index</td>
<td>867</td>
</tr>
</tbody>
</table>
Preface to the First Edition

The purpose of this book is for readers to learn how to apply statistical methods to the biomedical sciences. The book is written so that those with no prior training in statistics and a mathematical knowledge through algebra can follow the text—although the more mathematical training one has, the easier the learning. The book is written for people in a wide variety of biomedical fields, including (alphabetically) biologists, biostatisticians, dentists, epidemiologists, health services researchers, health administrators, nurses, and physicians. The text appears to have a daunting amount of material. Indeed, there is a great deal of material, but most students will not cover it all. Also, over 30% of the text is devoted to notes, problems, and references, so that there is not as much material as there seems to be at first sight. In addition to not covering entire chapters, the following are optional materials: asterisks (*) preceding a section number or problem denote more advanced material that the instructor may want to skip; the notes at the end of each chapter contain material for extending and enriching the primary material of the chapter, but this may be skipped.

Although the order of authorship may appear alphabetical, in fact it is random (we tossed a fair coin to determine the sequence) and the book is an equal collaborative effort of the authors. We have many people to thank. Our families have been helpful and long-suffering during the writing of the book: for LF, Ginny, Brad, and Laura; for GvB, Johanna, Loeske, William John, Gerard, Christine, Louis, and Bud and Stacy. The many students who were taught with various versions of portions of this material were very helpful. We are also grateful to the many collaborating investigators, who taught us much about science as well as the joys of collaborative research. Among those deserving thanks are for LF: Ed Alderman, Christer Allgulander, Fred Applebaum, Michele Battie, Tom Bigger, Stan Bigos, Jeff Borers, Martial Bourassa, Raleigh Bowden, Bob Bruce, Bernie Chaitman, Reg Clifft, Rollie Dickson, Kris Doney, Eric Foster, Bob Frye, Bernard Gersh, Karl Hammermeister, Dave Holmes, Mel Judkins, George Kaiser, Ward Kennedy, Tom Killip, Ray Lipicky, Paul Martin, George McDonald, Joel Meyers, Bill Myers, Michael Mock, Gene Passamani, Don Peterson, Bill Rogers, Tom Ryan, Jean Sanders, Lester Sauvage, Rainer Storb, Keith Sullivan, Bob Temple, Don Thomas, Don Weiner, Bob Witherspoon, and a large number of others. For GvB: Ralph Bradley, Richard Cornell, Polly Feigl, Pat Friel, Al Heyman, Myles Holland, Jim Hughes, Dave Kalman, Jane Koenig, Tom Koepsell, Bud Kukul, Eric Larson, Will Longstreth, Dave Luthy, Lorene Nelson, Don Martin, Duane Meeter, Gil Omenn, Don Peterson, Gordon Pledger, Richard Savage, Kirk Shy, Nancy Temkin, and many others. In addition, GvB acknowledges the secretarial and moral support of Sue Goleeke. There were many excellent and able typists over the years; special thanks to Myrna Kramer, Pat Coley, and Jan Alcorn. We owe special thanks to Amy Plummer for superb work in tracking down authors and publishers for permission to cite their work. We thank Robert Fisher for help with numerous figures. Rob Christ did an excellent job of using LATEX for the final version of the text. Finally, several people assisted with running particular examples and creating the tables; we thank Barry Storer, Margie Jones, and Gary Schoch.
Our initial contact with Wiley was the indefatigable Beatrice Shube. Her enthusiasm for our effort carried over to her successor, Kate Roach. The associate managing editor, Rose Ann Campise, was of great help during the final preparation of this manuscript.

With a work this size there are bound to be some errors, inaccuracies, and ambiguous statements. We would appreciate receiving your comments. We have set up a special electronic-mail account for your feedback:

http://www.biostat-text.info

LLOYD D. FISHER
GERALD VAN BELLE
Preface to the Second Edition

Biostatistics did not spring fully formed from the brow of R. A. Fisher, but evolved over many years. This process is continuing, although it may not be obvious from the outside. It has been ten years since the first edition of this book appeared (and rather longer since it was begun). Over this time, new areas of biostatistics have been developed and emphases and interpretations have changed.

The original authors, faced with the daunting task of updating a 1000-page text, decided to invite two colleagues to take the lead in this task. These colleagues, experts in longitudinal data analysis, survival analysis, computing, and all things modern and statistical, have given a twenty-first-century thrust to the book.

The author sequence for the first edition was determined by the toss of a coin (see the Preface to the First Edition). For the second edition it was decided to switch the sequence of the first two authors and add the new authors in alphabetical sequence.

This second edition adds a chapter on randomized trials and another on longitudinal data analysis. Substantial changes have been made in discussing robust statistics, model building, survival analysis, and discrimination. Notes have been added, throughout, and many graphs redrawn. We have tried to eliminate errata found in the first edition, and while more have undoubtedly been added, we hope there has been a net improvement. When you find mistakes we would appreciate hearing about them at http://www.vanbelle.org/biostatistics/.

Another major change over the past decade or so has been technological. Statistical software and the computers to run it have become much more widely available—many of the graphs and new analyses in this book were produced on a laptop that weighs only slightly more than a copy of the first edition—and the Internet provides ready access to information that used to be available only in university libraries. In order to accommodate the new sections and to attempt to keep up with future changes, we have shifted some material to a set of Web appendices. These may be found at http://www.biostat-text.info. The Web appendices include notes, data sets and sample analyses, links to other online resources, all but a bare minimum of the statistical tables from the first edition, and other material for which ink on paper is a less suitable medium.

These advances in technology have not solved the problem of deadlines, and we would particularly like to thank Steve Quigley at Wiley for his equanimity in the face of schedule slippage.

Gerald van Belle
Lloyd Fisher
Patrick Heagerty
Thomas Lumley

Seattle, June 15, 2003
CHAPTER 1

Introduction to Biostatistics

1.1 INTRODUCTION

We welcome the reader who wishes to learn biostatistics. In this chapter we introduce you to
the subject. We define statistics and biostatistics. Then examples are given where biostatistical
techniques are useful. These examples show that biostatistics is an important tool in advancing
our biological knowledge; biostatistics helps evaluate many life-and-death issues in medicine.

We urge you to read the examples carefully. Ask yourself, “what can be inferred from the
information presented?” How would you design a study or experiment to investigate the problem
at hand? What would you do with the data after they are collected? We want you to realize that
biostatistics is a tool that can be used to benefit you and society.

The chapter closes with a description of what you may accomplish through use of this book.
To paraphrase Pythagoras, there is no royal road to biostatistics. You need to be involved. You
need to work hard. You need to think. You need to analyze actual data. The end result will be
a tool that has immediate practical uses. As you thoughtfully consider the material presented
here, you will develop thought patterns that are useful in evaluating information in all areas of
your life.

1.2 WHAT IS THE FIELD OF STATISTICS?

Much of the joy and grief in life arises in situations that involve considerable uncertainty. Here
are a few such situations:

1. Parents of a child with a genetic defect consider whether or not they should have another
child. They will base their decision on the chance that the next child will have the same
defect.

2. To choose the best therapy, a physician must compare the prognosis, or future course, of
a patient under several therapies. A therapy may be a success, a failure, or somewhere
in between; the evaluation of the chance of each occurrence necessarily enters into the
decision.

3. In an experiment to investigate whether a food additive is carcinogenic (i.e., causes or at
least enhances the possibility of having cancer), the U.S. Food and Drug Administration
has animals treated with and without the additive. Often, cancer will develop in both the
treated and untreated groups of animals. In both groups there will be animals that do