Principles of
Biostatistics 2nd Ed

SECOND EDITION

Principles of Biostatistics

Marcello Pagano

Harvard School of Public Health

Kimberlee Gauvreau

Harvard Medical School

Duxbury
Thomson Learning™

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

Sponsoring Editor: *Carolyn Crockett*
Marketing Team: *Tom Ziolkowski, Beth Kroenke, Samantha Cabaluna*
Assistant Editor: *Seema Atwal*
Editorial Assistant: *Ann Day*
Production Editor: *Tom Novack*
Manuscript Editor: *Judith Abrahms*
Permissions Editor: *Connie Dowcett*
Interior Design: *John Edeen*
Cover Design: *Kelly Shoemaker*
Art Editor: *Jennifer Mackres*
Interior Illustration: *Pat Rogondino, and S. T. Associates Inc.*
Print Buyer: *Kris Waller*
Typesetting: *Omegatype Typography, Inc.*
Cover Printing: *R. R. Donnelley & Sons Co., Crawfordsville*
Printing and Binding: *R. R. Donnelley & Sons Co., Crawfordsville*

For more information about this or any other Duxbury products, contact:
DUXBURY
511 Forest Lodge Road
Pacific Grove, CA 93950 USA
www.duxbury.com
1-800-423-0563 (Thomson Learning Academic Resource Center)

Printed in United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

ISBN 0-534-22902-6

THIS BOOK IS PRINTED ON ACID-FREE RECYCLED PAPER

This book is dedicated with love to
Phyllis, John-Paul, Marisa, Loris, Alice and Lilian.
Neil and Eliza.

Preface

This book was written for students of the health sciençes and serves as an introduction to the study of biostatistics, or the use of numerical techniques to extract information from data and facts. Because numbers are more precise than words, they are particularly well suited for communicating scientific results.

However, just as one can lie with words, one can also lie with numbers. Indeed, numbers and lies have been linked for quite some time; there is even a book entitled *How to Lie with Statistics.* This association may owe its origin, or its affirmation at the very least, to the British prime minister Benjamin Disraeli. Disraeli is credited by Mark Twain as having said "There are three kinds of lies: lies, damned lies, and statistics." One has only to observe any modern political campaign to be convinced of the abuse of statistics. But enough about lies; this book adopts the position of Professor Frederick Mosteller, who said "It is easy to lie with statistics, but it is easier to lie without them."

Background

Principles of Biostatistics is aimed at students in the biological and health sciences who wish to learn modern research methods. It is based on a required course offered at the Harvard School of Public Health. In addition to these graduate students, a large number of health professionals from the Harvard medical area attend as well. The course is as old as the School itself, which attests to its importance. It spans 16 weeks of lectures and laboratory sessions. Each week includes two 50-minute lectures and one 2-hour lab. The entire class is together for the lectures, but is divided into smaller groups headed by teaching assistants for the lab sessions. These labs reinforce the material covered in the lectures, review the homework assignments, and introduce the computer into the course. We have included the lab materials—except those dealing with the homework assignments and specific computer commands—in the sections labeled Further Applications. These sections present either additional examples or a different perspective on the material covered in a chapter. They are designed to provoke discussion, although they are sufficiently complete for an individual who is not using the book as a course text to benefit from reading them.

This book has evolved to include topics that we believe can be covered at some depth in one American semester. Clearly, some choices had to be made; we hope that we have chosen well. In our course, we have sufficient time to cover most of the topics in the first 20 chapters. However, there is enough material presented to allow the instructor some flexibility. For example, some instructors may choose to omit the sections covering grouped data (Section 3.3), Chebychev's inequality (Section 3.4), and the Poisson distribution (Section 7.3), or the chapter on analysis of variance (Chapter 12), if they consider these concepts to be less important than others.

Structure

Some say that statistics is the study of variability and uncertainty. We believe there is truth to this adage, and have used it as a guide in dividing the book into three parts. The first five chapters deal with collections of numbers and ways in which to summarize, explore, and explain them. The next two chapters focus on probability and serve as an introduction to the tools needed for the subsequent investigation of uncertainty. It is only in the eighth chapter and thereafter that we distinguish between populations and samples and begin to investigate the inherent variability introduced by sampling, thus progressing to inference. We think that this modular introduction to the quantification of uncertainty is justified by the success achieved by our students. Postponing the slightly more difficult concepts until a solid foundation has been established makes it easier for the reader to comprehend them.

Data Sets and Examples

Throughout the text we have used data drawn from published studies to exemplify biostatistical concepts. Not only is real data more meaningful, it is usually more interesting as well. Of course, we do not wish to use examples in which the subject matter is too esoteric or too complex. To this end, we have been guided by the backgrounds and interests of our students—primarily topics in public health and clinical research—to choose examples that best illustrate the concepts at hand.

There is some risk involved in using published data. We cannot guarantee that all of the examples are honest and that the data were properly collected; for this we must rely on the reputations of our sources. We do not belittle the importance of this consideration. The value of our inference depends critically on the worth of the data, and we strongly recommend that a good deal of effort be expended on evaluating its quality. We assume that this is understood by the reader.

More than once we have used examples in which the population of the United States is broken down along racial lines. In reporting these official statistics we follow the lead of the government agencies that release them. We do not wish to reify this racial categorization, since in fact the observed differences may well be due to socioeconomic factors rather than the implied racial ones. One option would be to ignore these statistics; however, this would hide inequities which exist in our health system—inequities that need to be eliminated. We focus attention on the problem in the hope of stimulating interest in promoting solutions.

We have minimized the use of mathematical notation because of its well-deserved reputation of being the ultimate jargon. If used excessively, it can intimidate even the most ardent scholar. We do not wish to eliminate it entirely, however; it has been developed over the ages to be helpful in communicating results. We hope that in this respect we have written a succinct and understandable text.

Over and above their precision, there is something more to numbers—maybe a little magic—that makes them fun to study. The fun is in the conceptualization more than the calculations, and we are fortunate that we have the computer to do the drudge work. This allows students to concentrate on the ideas. In other words, the computer allows the instructor to teach the poetry of statistics and not the plumbing.

Computing

To take advantage of the computer, one needs a good statistical package. We use Stata, which is available from the Stata Corporation in College Station, Texas. We find this statistical package to be one of the best on the market today; it is user-friendly, accurate, powerful, reasonably priced, and works on a number of different platforms, including Windows, Unix, and Macintosh. Furthermore, the output from this package is acceptable to the Federal Drug Administration in New Drug Approval submissions. Other packages are available, and this book can be supplemented by any one of them. In this second edition, we also present output from SAS and Minitab in the Further Applications section of each chapter. We strongly recommend that some statistical package be used.

Some of the review exercises in the text require the use of the computer. To help the reader, we have included the data sets used in these exercises both in Appendix B and on a CD at the back of the book. The CD contains each data set in two different formats: an ASCII file (the "raw" suffix) and a Stata file (the "dta" suffix). There are also many exercises that do not require the computer. As always, active learning yields better results than passive observation. To this end, we cannot stress enough the importance of the review exercises, and urge the reader to attempt as many as time permits.

New to the Second Edition

This second edition includes revised and expanded discussions on many topics throughout the book, and additional figures to help clarify concepts. Previously used data sets, especially official statistics reported by government agencies, have been updated whenever possible. Many new data sets and examples have been included; data sets described in the text are now contained on the CD enclosed with the book. Tables containing exact probabilities for the binomial and Poisson distributions (generated by Stata) have been added to Appendix A. As previously mentioned, we now incorporate computer output from SAS and Minitab as well as Stata in the Further Applications sections. We have also added numerous new exercises, including questions reviewing the basic concepts covered in each chapter.

Acknowledgements

A debt of gratitude is owed a number of people: Harvard University President Derek Bok for providing the support which got this book off the ground, Dr. Michael K. Martin for calculating Tables A.3 through A.8 in Appendix A, and John-Paul Pagano for

assisting in the editing of the first edition. We thank the individuals who reviewed the manuscript: Rick Chappell, University of Wisconsin; Dr. Todd G. Nick, University of Mississippi Medical Center; Al Bartolucci, University of Alabama at Birmingham; Bruce E. Trumbo, California State University, Hayward; James Godbold, The Mount Sinai School of Medicine of New York University; and Maureen Lahiff, University of California, Berkeley. Our thanks to the teaching assistants who have helped us teach the course and who have made many valuable suggestions. Probably the most deserving of thanks are the students who have taken the course over the years and who have tolerated us as we learned how to teach it. We are still learning.

Marcello Pagano
Kimberlee Gauvreau

Boston, Massachusetts

Contents

Introduction

In 1903, H. G. Wells hypothesized that statistical thinking would one day be as necessary for good citizenship as the ability to read and write. Statistics do play an important role in many decision-making processes. Before a new drug can be marketed, for instance, the United States Food and Drug Administration requires that it be subjected to a clinical trial, an experimental study involving human subjects. The data from this study must be compiled and analyzed to determine whether the drug is not only effective, but safe. In addition, the U.S. government's decisions regarding Social Security and public health programs rely in part on predictions about the longevity of the nation's population; consequently, it must be able to predict the number of years that each individual will live. Many other issues need to be addressed as well. Where should a government invest its resources if it wishes to reduce infant mortality? Does the use of a seat belt or an air bag decrease the chance of death in a motor vehicle accident? Should a mastectomy always be recommended to a patient with breast cancer? What factors increase the risk that an individual will develop coronary heart disease? To answer these questions and others, we rely on the methods of biostatistics.

The study of *statistics* explores the collection, organization, analysis, and interpretation of numerical data. The concepts of statistics may be applied to a number of fields that include business, psychology, and agriculture. When the focus is on the biological and health sciences, we use the term *biostatistics*.

Historically, statistics have been used to tell a story with numbers. Numbers often communicate ideas more succinctly than do words. The message carried by the following data is quite clear, for instance. In 1979, 48 persons in Japan, 34 in Switzerland, 52 in Canada, 58 in Israel, 21 in Sweden, 42 in Germany, 8 in England, and 10,728 in the United States were killed by handguns [1]. The power of these numbers is obvious; the point would be made even if we were to correct for differences in population size.

As a second example, consider the following quotation, taken from an editorial in *The Boston Globe* [2]:

> Lack of contraception is linked to an exceptionally high abortion rate in the Soviet Union—120 abortions for every 100 births, compared with 20 per 100 births in