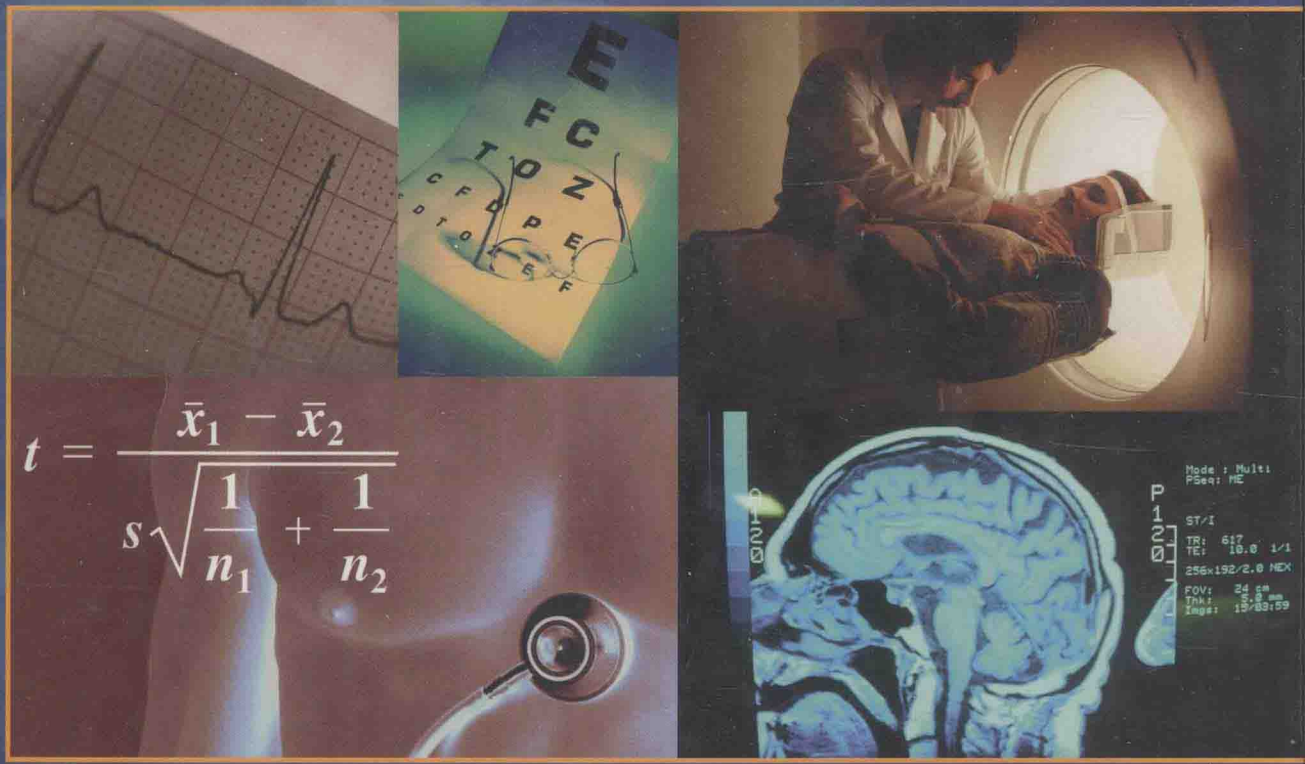
5TH EDITION

FUNDAMENTALS OF BIOSTATISTICS

BERNARD ROSNER



Fifth Edition

Fundamentals of Biostatistics

Bernard Rosner

Harvard University

Duxbury
Thomson Learning™

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

Sponsoring Editor: *Carolyn Crockett*
Marketing Team: *Samantha Cabaluna, Beth Kronke, and Tom Ziolkowski*
Editorial Assistants: *Ann Day and Kimberly Raburn*
Production Editor: *Mary Vezilich*
Manuscript Editor: *Linda Purrington*
Permissions Editor: *Lillian Campobasso*
Production Service: *Scratchgravel Publishing Services*

Cover Design: *Laurie Albrecht*
Cover Photos: *PhotoDisc*
Interior Design: *Devenish Design*
Art Editor: *Lisa Torri*
Art Illustration: *Suffolk Graphics*
Print Buyer: *Vena Dyer*
Printing and Binding: *R. R. Donnelley/Crawfordsville*

For more information, contact:

DUXBURY
511 Forest Lodge Road
Pacific Grove, CA 93950 USA
www.duxbury.com

Printed in the United States of America

10 9 8 7 6

Library of Congress Cataloging-in-Publication Data
Rosner, Bernard (Bernard A.)
Fundamentals of biostatistics / Bernard Rosner.—5th ed.
p. cm.
Includes bibliographical references and indexes.
ISBN 0-534-37068-3 (alk. paper)
1. Biometry. 2. Medical statistics. I. Title.

QH323.5 .R674 2000
570'.1'5195—dc21 99-050394

Fifth Edition

Fundamentals of Biostatistics

Duxbury Titles of Related Interest

Carver, *Doing Data Analysis with MINITAB 12*

Elliott, *Learning SAS in the Computer Lab,* 2nd ed.

Johnson, *Applied Multivariate Methods for Data Analysis*

Kleinbaum/Kupper/Muller/Nizam, *Applied Regression Analysis and Multivariable Methods,* 3rd ed.

Kuehl, *Design of Experiments: Statistical Principles of Research Design and Analysis,* 2nd ed.

Lehmann/Zeitz, *Statistical Explorations with Microsoft Excel*

Lohr, *Sampling: Design and Analysis*

MathSoft Inc., *S-Plus 4.5 for Windows, Student Edition*

SAS Institute Inc., *JMP-In: Statistical Discovery Software*

Selvin, *Practical Biostatistical Methods*

This book is dedicated to my wife, Cynthia, and my children, Sarah, David, and Laura

Preface

This introductory-level biostatistics text is designed for upper-level undergraduate or graduate students interested in medicine or other health-related areas. It requires no previous background in statistics, and its mathematical level assumes only a knowledge of algebra.

Fundamentals of Biostatistics evolved from notes that I have used in a biostatistics course taught to Harvard University undergraduates and Harvard Medical School students over the past twenty-five years. I wrote this book to help motivate students to master the statistical methods that are most often used in the medical literature. From the student's viewpoint, it is important that the example material used to develop these methods is representative of what actually exists in the literature. Therefore, most of the examples and exercises in this book are based either on actual articles from the medical literature or on actual medical research problems I have encountered during my consulting experience at the Harvard Medical School.

The Approach

Most introductory statistics texts either use a completely nonmathematical, cookbook approach or develop the material in a rigorous, sophisticated mathematical framework. In this book, however, I follow an intermediate course, minimizing the amount of mathematical formulation but giving complete explanations of all the important concepts. Every new concept in this book is developed systematically through completely worked-out examples from current medical research problems. In addition, I introduce computer output where appropriate to illustrate these concepts.

I initially wrote this text for the introductory biostatistics course. However, the field has changed rapidly over the past several years; because of the increased power of newer statistical packages, we can now perform more sophisticated data analyses than ever before. Therefore, a second goal of this text is to present these new techniques *at an introductory level* so that students can become familiar with them without having to wade through specialized (and, usually, more advanced) statistical texts.

To differentiate these two goals more clearly, I included most of the content for the introductory course in the first 12 chapters. I then added a new chapter (Chapter 13), "Design and Analysis Techniques for Epidemiologic Studies." This new chapter, together with Chapter 14, "Hypothesis Testing: Person-Time Data," covers more advanced statistical techniques used in recent epidemiologic studies.

Changes in the Fifth Edition

For this edition, I have added 11 new sections and substantially revised eight other sections. Features new to this edition include the following:

- An expanded set of computer exercises based on real data sets. This edition contains 23 data sets, which are contained on the disk bound in the back of the book. For the first time you will find each data set available in the Excel-readable format, in addition to the MINITAB® and ASCII formats found in the previous edition.
- An additional case study—concerning the effects of cigarette smoking on bone density—used in several chapters throughout the book.
- New or expanded coverage of the following topics:
  - ROC curves (Section 3.7.1)
  - Use of electronic tables based on Microsoft® Excel (Section 4.8.2)
  - Covariance (Sections 5.6.1 and 11.7)
  - Expected value and variance of linear combinations of dependent random variables (Section 5.6.1)
  - Use of Excel to perform hypothesis tests and obtain confidence intervals (Chapters 6–14)
  - Sample size estimation for longitudinal studies (Section 8.11)
  - The delta method (Section 13.3)
  - Equivalence studies (Section 13.9)
  - Meta-analysis (Section 13.8)
  - Variance-covariance matrix (Section 12.5)
  - Sample size estimation for correlation coefficients (Section 11.8.4)
  - Clustered binary data (Section 13.11)
  - Measurement error methods (Section 13.12)
  - Power and sample size estimation for person-time data (Sections 14.4 and 14.6)
  - Sample size estimation for survival analysis (Section 14.12)
  - One-sample inference for incidence rate data (Section 14.2)
  - Confidence intervals for incidence rates (Section 14.2)

The new sections and the expanded sections for this edition have been indicated by an asterisk in the table of contents.

Exercises

This edition contains 1166 exercises (compared with 893 in the previous edition). All data-set-based problems are included. Problems marked by an asterisk (*) at the end of each chapter have corresponding brief solutions in the answer section at the back of the book. Based on requests from students for more completely solved problems, approximately 600 additional problems are presented in the *Study Guide to Accompany Fundamentals of Biostatistics,* 5th edition (ISBN 0–534–37120–5). The study guide presents a complete solution for each of the problems it contains. In addition, approximately 100 of these problems are included in a Miscellaneous Problems section and are randomly ordered so that they are not tied to a specific chapter in the book. This gives the student additional practice in determining what method to use in what situation. Finally, the appendix to the *Study Guide* has a brief description of the statistical commands in Excel used in this textbook.

Computation Method

The method of handling computations is similar to the fourth edition. All intermediate results are carried to full precision (10+ significant digits), even though they are presented with fewer significant digits (usually 2 or 3) in the text. Thus, intermediate results may seem inconsistent with final results in some instances; this, however, is not the case.

Organization

Fundamentals of Biostatistics, 5th edition, is organized as follows.

Chapter 1 is an *introductory* chapter that contains an outline of the development of an actual medical study with which I was involved. It provides a unique sense of the role of biostatistics in medical research.

Chapter 2 concerns *descriptive statistics* and presents all the major numeric and graphic tools used for displaying medical data. This chapter is especially important for both consumers and producers of medical literature because much actual communication of information is accomplished via descriptive material.

Chapters 3 through 5 discuss *probability*. The basic principles of probability are developed, and the most common probability distributions—such as the binomial and normal distributions—are introduced. These distributions are used extensively in later chapters of the book.

Chapters 6 through 10 cover some of the basic methods of *statistical inference.*

Chapter 6 introduces the concept of drawing random samples from populations. The difficult notion of a sampling distribution is developed and includes an introduction to the most common sampling distributions, such as the t and chi-square distributions. The basic methods of *estimation,* including an extensive discussion of confidence intervals, are also presented.

Chapters 7 and 8 contain the basic principles of *hypothesis testing.* The most elementary hypothesis tests for normally distributed data, such as the t test, are also fully discussed for one- and two-sample problems.

Chapter 9 covers the basic principles of *nonparametric statistics*. The assumptions of normality are relaxed, and distribution-free analogues are developed for the tests in Chapters 7 and 8.

Chapter 10 contains the basic concepts of *hypothesis testing* as applied to categorical data, including some of the most widely used statistical procedures, such as the chi-square test and Fisher's exact test.

Chapter 11 develops the principles of *regression analysis*. The case of simple linear regression is thoroughly covered, and extensions are provided for the multiple regression case. Important sections on goodness-of-fit of regression models are also included. Finally, rank correlation is introduced.

Chapter 12 introduces the basic principles of the *analysis of variance* (ANOVA). The one-way analysis of variance fixed and random effects models are discussed. In addition, two-way ANOVA and the analysis of covariance are covered. Finally, we discuss nonparametric approaches to one-way ANOVA.

Chapter 13 discusses methods of design and analysis for *epidemiologic studies*. The most important study designs, including the prospective study, the case–control study, the cross-sectional study, and the cross-over design are introduced. The concept of a confounding variable—that is, a variable related to both the disease and the exposure variable—is introduced, and methods for controlling for confounding, which include the Mantel-Haenszel test and multiple-logistic regression, are discussed in detail. This discussion is followed by the exploration of topics of current interest in epidemiologic data analysis, including: meta-analysis (the combination of results from more than one study); correlated binary data techniques (techniques that can be applied when replicate measures, such as data from multiple teeth from the same person, are available for an individual); measurement error methods (useful when there is substantial measurement error in the exposure data collected); and equivalence studies (whose objective it is to establish bioequivalence between two treatment modalities rather than that one treatment is superior to the other).

Chapter 14 introduces methods of analysis for person-time data. The methods covered in this chapter include those for incidence-rate data, as well as several methods of survival analysis: the Kaplan-Meier survival curve estimator, the log rank test, and the Cox proportional hazards model.

Throughout the text—particularly in Chapter 13—I discuss the elements of study designs, including the concepts of matching; cohort studies; case–control studies; retrospective studies; prospective studies; and the sensitivity, specificity, and predictive value of screening tests. These designs are presented in the context of actual samples. In addition, Chapters 7, 8, 10, 11, 13, and 14 contain specific sections on sample-size estimation for different statistical situations.

A flowchart of appropriate methods of statistical inference (see pages 776–780) is a handy reference guide to the methods developed in this book. At the end of each of Chapters 7 through 14, I refer students to this flowchart in order to give them some perspective on how the methods discussed in a given chapter fit in with all the other statistical methods introduced in this book.

In addition, I have provided an index of applications, grouped by *medical specialty*, that summarizes all the examples and problems that this book covers.

Acknowledgments

I am indebted to Debra Sheldon, the late Marie Sheehan, and Harry Taplin for their invaluable help typing the manuscript, and to Marion McPhee for helping to prepare the disk. I am also indebted to the manuscript reviewers, among them: John E. Alcaraz, San Diego State University; Stewart J. Anderson, the University of Pittsburgh; Christiana Drake, the University of California—Davis; P. D. M. Macdonald, McMaster University; Craig D. Turnbull, the University of North Carolina at Chapel Hill; Mark J. van der Laan, the University of California—Berkeley; and Dennis Wallace, the University of Kansas School of Medicine. I would also like to thank my colleagues Nancy Cook, Robert Glynn, Cathy Berkey, and Donna Spiegelman for their helpful reviews of new material in this edition.

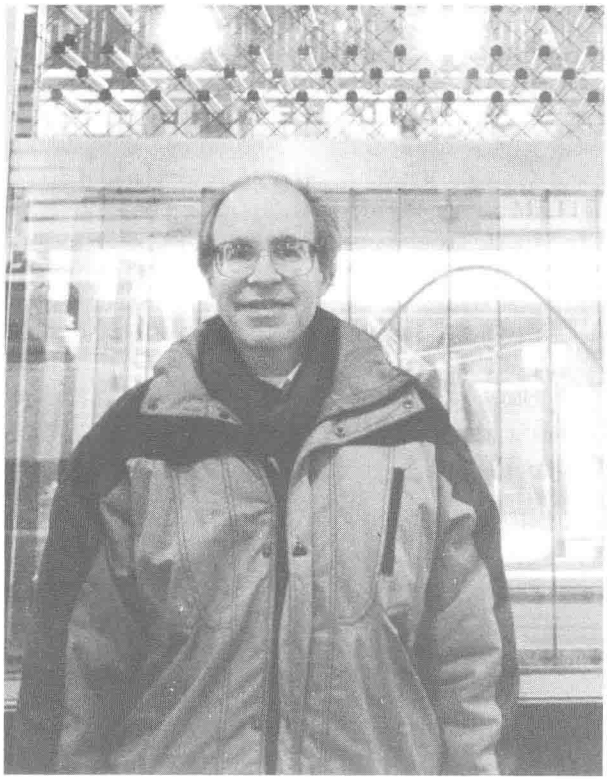
In addition, I wish to thank Carolyn Crockett, Mary Vezilich, Anne Draus, Greg Draus, and Linda Purrington, who were instrumental in providing editorial advice and in preparing the manuscript.

I am also indebted to my colleagues at the Channing Laboratory—most notably, the late Edward Kass, Frank Speizer, Charles Hennekens, the late Frank Polk, Ira Tager, Jerome Klein, James Taylor, Stephen Zinner, Scott Weiss, Frank Sacks, Walter Willett, Alvaro Munoz, Graham Colditz, and Susan Hankinson—and to my other colleagues at the Harvard Medical School—most notably, Frederick Mosteller, Eliot Berson, Robert Ackerman, Mark Abelson, Arthur Garvey, Leo Chylack, Eugene Braunwald, and Arthur Dempster, who inspired me to write this book. I also wish to acknowledge John Hopper and Philip Landrigan for providing the data for our case studies.

Finally, I would like to acknowledge Leslie Miller, Andrea Wagner, Loren Fishman, and Frank Santopietro, without whose clinical help the current edition of this book would not have been possible.

Bernard Rosner

About the Author



Bernard Rosner is Professor of Medicine (Biostatistics) at Harvard Medical School and Professor of Biostatistics in the Harvard School of Public Health. He received a B.A. in Mathematics from Columbia University in 1967, an M.S. in Statistics from Stanford University in 1968, and a Ph.D. in Statistics from Harvard University in 1971.

He has more than 25 years of biostatistical consulting experience with other investigators at the Harvard Medical School. Special areas of interest include: cardiovascular disease, hypertension, breast cancer, and ophthalmology. Many of the examples and exercises used in the text reflect data collected from actual studies in conjunction with his consulting experience. In addition, he has developed new biostatistical methods, mainly in the areas of longitudinal data analysis, analysis of clustered data (such as data collected in families or from paired organ systems in the same person), measurement error methods, and outlier detection methods. You will see some of these methods introduced in this book at an elementary level. He was married in 1972 to his wife, Cynthia, and has three children, Sarah, David, and Laura, each of whom has contributed examples for this book.

Contents

*The new sections and the expanded sections for this edition are indicated by an asterisk.

CHAPTER 4

Discrete Probability Distributions / 79

CHAPTER 5

Continuous Probability Distributions / 117

CHAPTER 6

Estimation / 157

CHAPTER 7

Hypothesis Testing: One-Sample Inference / 211

CHAPTER 8

Hypothesis Testing: Two-Sample Inference / 273

CHAPTER 9

Nonparametric Methods / 331

CHAPTER 10

Hypothesis Testing: Categorical Data / 355

CHAPTER 11

Regression and Correlation Methods / 425

CHAPTER 12

Multisample Inference / 511