

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, Harvard Medical School, and Harvard School of Public Health students over the past 30 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 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 dramatically over the past 30 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. More advanced statistical techniques used in recent epidemiologic studies are covered in Chapter 13, "Design and Analysis Techniques for Epidemiologic Studies," and Chapter 14, "Hypothesis Testing: Person-Time Data."

Changes in the Eighth Edition

For this edition, I have added three new sections and added new content to three other sections. Features new to this edition include the following:

- The data sets are now available on the book's Companion Website at www.cengage.com/statistics/rosner in an expanded set of formats, including Excel, Minitab®, SPSS, JMP, SAS, Stata, R, and ASCII formats.
- Data and medical research findings in Examples have been updated.
- New or expanded coverage of the followings topics has been added:
 - The Bootstrap (Section 6.11)
 - One-sample inference for the Binomial Distribution (Section 7.9)
 - Permutation Tests (Section 9.6)
 - Sample size estimation for logistic regression (Section 13.9)
 - Estimation of survival curves: The Kaplan-Meier Estimator (Section 14.9)
 - Derivation of selected formulas (Sections 7.12, 8.11, 10.9, 11.14, 12.11, 13.17, 14.15)

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 1,490 exercises; 171 of these exercises are new. Data and medical research findings in the problems have been updated where appropriate. All problems based on the data sets 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 and complete solutions are presented in the *Study Guide* available on the Companion Website accompanying this text. 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. Complete instructor solutions to all exercises are available at the instructor companion website at cengage.com/statistics/rosner.

Computation Method

The method of handling computations is similar to that used in the seventh 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, Eighth 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 information is actually communicated 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. The concepts of prior probability and posterior probability are also introduced.

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. *In addition, the bootstrap method for obtaining confidence limits is introduced for the first time.*

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. *The technique of permutation testing, which is widely used in genetic studies, is introduced for the first time.*

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. Also, rank correlation is introduced, including methods for obtaining confidence intervals for rank correlation.

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, the analysis of covariance, and mixed effects models are covered. Finally, we discuss nonparametric approaches to one-way ANOVA. Multiple comparison methods including material on the false discovery rate are also provided.

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. Extensions to logistic regression models, including conditional logistic regression, polytomous logistic regression, and ordinal logistic regression, are discussed. *Methods of estimation of sample size for logistic regression models are provided for the first time.* 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); equivalence studies (whose objective it is to establish bioequivalence between two

treatment modalities rather than that one treatment is superior to the other); and missing-data methods for how to handle missing data in epidemiologic studies. Longitudinal data analysis and generalized estimating equation (GEE) methods are also briefly discussed.

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 proportional-hazards model. Methods for testing the assumptions of the proportional-hazards model have also been included. Parametric survival analysis methods are also discussed.

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.

There have been two important organizational changes in the presentation of material in the text. First, the derivation of more complex formulas have either been moved after the statement of an equation or to separate derivation sections at the end of the chapter, to enable students to access the main results in the equations more immediately. Second, there are numerous subsections entitled “Using the Computer to Perform a Specific Test” to more clearly highlight use of the computer to implement many of the methods in the text.

A flowchart of appropriate methods of statistical inference (see pages 895–900) is a handy reference guide to the methods developed in this book. Page references for each major method presented in the text are also provided. In Chapters 7 and 8 and Chapters 10–14, I refer students to this flowchart to give them some perspective on how the methods discussed in a given chapter fit with all the other statistical methods introduced in this book.

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

Finally, we provide for the first time, an index of computer software to more clearly identify the computer commands in specific computer packages that are featured in the text.

Acknowledgments

I am indebted to Debra Sheldon, the late Marie Sheehan, and Harry Taplin for their invaluable help typing the manuscript, to Dale Rinkel for invaluable help in typing problem solutions, and to Marion McPhee for helping to prepare the data sets on the Companion Website. I am also indebted to Roland Matsouaka for updating solutions to problems for this edition, and to Virginia Piaseczny for typing the Index of Applications. In addition, I wish to thank the manuscript reviewers, among them: Shouhao Zhou, Daniela Szatmari-Voicu, Jianying Gu, Raid Amin, Claus Wilke, Glen Johnson, Kara Zografos, and Hui Zhao. I would also like to thank my colleagues Nancy Cook, who was instrumental in helping me develop the part of Section 12.4 on the false-discovery rate, and Robert Glynn, who was invaluable in developing Section 13.16 on missing data and Section 14.11 on testing the assumptions of the proportional-hazards model.

In addition, I wish to thank Spencer Arritt and Jay Campbell, whose input was critical 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 Muñoz, Graham Colditz, and Susan Hankinson—and to my other colleagues at the Harvard Medical School, most notably, the late 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 express appreciation to John Hopper and Philip Landrigan for providing the data for our case studies.

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

Bernard Rosner