PREFACE

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

Fundamentals of Biostatistics evolved from a set of notes that I used in a course in biostatistics taught to Harvard University undergraduates and Harvard Medical School students over the past fifteen 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 examples and exercises used in this book are either based 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 other introductory statistics texts either use a completely nonmathematical, cookbook approach or develop the material in a rigorous, sophisticated mathematical framework. In this book I have attempted to follow an intermediate course, minimizing the amount of mathematical formulation and yet giving complete explanations of all the important concepts. *Every* new concept is developed systematically through completely worked out examples from current medical research problems. In addition, computer output is introduced where appropriate to illustrate these concepts.

The material in this book is suitable for either a one- or two-semester course in biostatistics. The material in Chapters 1 through 8 and Chapter 10 is suitable for a one-semester course. The instructor may select appropriate material from the other chapters as time permits.

Changes in the Fourth Edition

There are a total of 21 new sections and 9 additional sections with substantial revisions in the Fourth Edition. The new features include:

- An expanded set of computer exercises based on real data sets has been developed. The data sets are on a diskette that is provided with the book.
- A case study on lead exposure in children used in several chapters throughout the book.
- An extended discussion of randomized clinical trials including
  (a) design features (Section 6.4.1)
  (b) sample size issues (Section 10.7.3)
- One-Sample Inference for the Poisson distribution (Section 6.8 and 7.11)
- Sample size estimation based on confidence interval width (Section 7.7.3)
- Outlier detection techniques (Section 8.9)
- The one-way ANOVA random effects model (Section 9.6)

- The cross-over design (Section 9.7)
- A discussion of the most popular study designs in biomedical research (Section 10.3)
- Measures of effect for categorical data (Section 10.4)
- The hypergeometric distribution (Section 10.5.1)
- Issues in epidemiologic research include confounding, standardization, and effect modification (Sections 10.9 and 10.10)
- The Mantel extension test (Section 10.10.4)
- Power and sample size estimation for stratified categorical data (Section 10.11)
- Greatly expanded section on assessing goodness of fit of regression models including residual analysis for both simple linear regression (Section 11.6) and multiple linear regression (Section 11.7.3)
- Partial regression coefficients (Section 11.7.1)
- Partial residual plots (Section 11.7.3)
- Relationship between t test methods, analysis of variance, analysis of covariance, and regression analysis (Section 11.8)
- Interval estimation for correlation coefficients (Section 11.11.3)
- Partial and multiple correlation (Section 11.12)
- Intraclass correlation coefficient (Section 11.13)
- Expanded discussion of multiple logistic regression including methods for prediction and assessment of goodness of fit (Sections 11.14.4 and 11.14.5)
- A new chapter on inference for person-time data (Chapter 13), including
- Measures of effect for person-time data (Section 13.1)
- Inference for stratified person-time data (Section 13.3)
- Power and sample size estimation for person-time data (Section 13.4)
- Testing for trend with incidence rate data (Section 13.5)
- Estimation of survival curves with the Kaplan-Meier estimator (Section 13.7)

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

The Exercises

There are a total of 1600 exercises in the Fourth Edition (compared with 1300 in the Third Edition). Students have indicated that they would like to see more completely solved problems. As a result, 639 of the problems have been moved to a Study Guide to accompany the text given with complete solutions. 95 of these problems are given 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." The remaining 544 problems are related to specific chapters in the text. All problems have complete solutions. Approximately 900 problems remain in the text, including all data-set based problems. Brief solutions are given to 300 of these problems in the Answer section, and are indicated by an asterisk (*) in the problem section of each chapter.

A Method of Computation

The method of handling computations in this edition of the book has also changed. All intermediate results are carried to full precision (10+ significant digits) even though they are presented with fewer significant digits (usually 2–3) in the text. Thus, intermediate results may seem to be inconsistent with final results in some instances, although this is not the case. This method allows for greater accuracy of final results and is the reason why there are slightly different results given for many calculations versus the previous editions, where intermediate results were carried to the same precision as shown in the text.

Organization

Fundamentals of Biostatistics, fourth edition, is organized as follows:

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

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, since much of the 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 the 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 also developed, including an introduction to the most common sampling distributions, such as the t and chi-square distributions. The basic methods of *estimation* are also presented, including an extensive discussion of confidence intervals.

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 introduces the basic principles of the *analysis of variance* (ANOVA). The one-way analysis of variance fixed and random effects models are discussed as well as the analysis of data obtained using crossover designs.

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. Multiple logistic regression is also discussed.

Chapter 12 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, 8, 9, and 11.

Chapter 13 introduces methods of analysis for person-time data. Included are methods for incidence rate data, as well as methods of survival analysis including the Kaplan-Meier survival curve estimator, the log rank test, and the Cox proportional hazards model.

The elements of study design are also discussed, 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, specific sections on sample size estimation are provided for different statistical situations in Chapters 7, 8, 9, and 10.

A flowchart of appropriate methods of statistical inference on pages 671–675 provides an easy reference to the methods developed in this book. This flowchart is referred to at the end of each of Chapters 7 through 13 to give the student some perspective on how the methods in a particular chapter fit in with the overall collection of statistical methods introduced in this book.

In addition, an index summarizing all examples and problems used in this book is provided, grouped by *medical specialty*.

Acknowledgments

I am indebted to Debra Sheldon, Marie Sheehan, and Harry Taplin, who have been invaluable in helping to type this manuscript. I am indebted to those who reviewed the manuscript, among them: Stuart J. Anderson, University of Pittsburgh; Kenneth J. Koehler, Iowa State University; Donald J. Slymen, San Diego State University; and Craig D. Turnbull, University of North Carolina, Chapel Hill. I wish to thank Alex Kugushev, Jennie Burger, George Calmenson, and Linda Purrington, who were instrumental in providing editorial advice and in the preparation of the manuscript. I am indebted to my many colleagues at the Channing Laboratory, most notably Edward Kass, Frank Speizer, Charles Hennekens, 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 provided the inspiration for writing this book. Finally, I wish to acknowledge Leslie Miller, Andrea Wagner, Loren Fishman, and Roberta Shapiro, without whose clinical help the current edition of this book would not have been possible.

Bernard Rosner
Boston

CONTENTS

*asterisks indicate new section or subsection for the Fourth Edition

CHAPTER 8 Hypothesis Testing: Two-Sample Inference 251

CHAPTER 9 Multisample Inference 299

CHAPTER 10 Hypothesis Testing: Categorical Data 345