

PREFACE

As the title indicates, this book is intended to be a practical guide for the statistical analysis of biomarker data. To us, such a guide should include information on the proper application of statistical methods that are most commonly used to analyze biomarker data, with special emphasis placed on the underlying assumptions. This includes recommendations concerning: (1) preferred methods for determining whether or not the underlying assumptions are valid for a particular set of biomarker data and (2) how to proceed if the underlying assumptions appear to be violated. In addition to emphasizing the underlying assumptions, we have also placed considerable emphasis on computational issues related to the methods most commonly used to analyze biomarker data. To the greatest extent possible, we have provided software (primarily SAS[®] code) for performing the statistical methods that we recommend. All of this software, along with the complete data sets for all of the examples, is included in the Software Appendix that is available on the companion website for this textbook. For many of the examples, the complete data set, along with the computer code for analyzing the data, are also provided in the text.

Our intention has been to present our descriptions of the statistical methods that are included in this textbook at a fairly low technical level; however, we have provided all of the necessary formulas for those who are interested in the more theoretical aspects of these methods. By including these formulas, we hoped to achieve two goals: (1) to provide sufficient information for those who wish to program these methods themselves, using Excel, R or C++, for example and (2) to provide the mathematical details for those who are interested in the theoretical foundations of these methods. Sections that we feel are unusually technical (in terms of extensive use of formulas and/or theoretical development) are marked with an asterisk (*) and can be safely skipped without loss of continuity. Overall, the presentation in this textbook

does not require familiarity with calculus or matrix algebra. Generally speaking, a sufficient background for the material in this book would be the equivalent of a two-semester statistical methods sequence at the advanced undergraduate or introductory graduate level. Minimal requirements include a good working knowledge of basic statistical inference (point and interval estimation and hypothesis testing), as well as some familiarity with correlation analysis, chi-square analysis, and nonparametric statistics. Chapter 3 contains a review of the basic statistical concepts and methods that we consider to be sufficient prerequisites for the material in Chapters 4 and 5.

One may question our choice of statistical methods included in this book. For example, there is very little discussion of multivariable models, although we do provide a very elementary description of multiple linear regression in Chapter 3. We also briefly mention multiple logistic regression in several places. We realize the importance of the application of these and other multivariable methods in biomarker research (e.g., in the analysis of survival data or longitudinal data), but feel that we cannot fully address the numerous complex considerations involved in the appropriate use of these methods within the limited scope of this book. There is very heavy emphasis (some may say too heavy) in this book on testing distributional assumptions, and on nonparametric and exact methods. As illustrated in Section 4.2, investigators often state that they used a data transformation or nonparametric method to analyze their biomarker data due to "extreme skewness" (or similar language); however, very rarely do they provide any statistical justification for following this course of action. Many of the studies involving biomarker data that we have worked on as collaborating statisticians have had very small sample sizes (some as small as 5 or 6), hence the strong emphasis on exact methods. Another area that some may say has received too much emphasis in this book is the assessment of the reliability and validity of a biomarker. We feel that this aspect of biomarker development is too often ignored (or given only minimal attention) in the biomarker literature. Thus, one of our goals in writing this textbook has been to provide sufficient background knowledge and software tools so that it will be easier for investigators to undertake a more thorough assessment of these important characteristics of a newly proposed biomarker.

Wherever possible, we have illustrated the statistical methods included in this textbook with examples that make use of "real" data. Most of these data sets were taken (or adapted) from the biomarker literature, or from unpublished studies we participated in as collaborators. In some instances, we have used hypothetical data (which we hope are also realistic) when the "real data," we needed to illustrate a particular point were not readily available. Most of these hypothetical data sets were constructed so that they retained certain key features of the published study on which they were based.

We have not written this book with the anticipation that it would be used as the primary textbook for a course on the statistical analysis of biomarker data; however, it could be used for that purpose for a summer course, an independent study, or a graduate seminar. With this in mind, we have included a small number of problems (and solutions) at the end of each chapter. We consider our primary audience for this textbook to consist of investigators who wish to perform their own analyses of

biomarker data; this includes toxicologists, pharmacologists, epidemiologists, environmental and clinical laboratory scientists, and other professionals in the health and environmental sciences. Statisticians who routinely analyze and interpret biomarker data could also find it to be of interest. Since most of the statistical techniques that we discuss in this book also apply to surrogate markers (or, more accurately, surrogate endpoints), professionals who work in the areas of pharmaceutical and chemical product development could also find this book to be useful. Our goal has been to provide enough practical information (including software code and software recommendations) to facilitate the statistical analysis of biomarker data for all of these individuals.