Preface

Biostatistics applies the principles of statistics to the medical or health fields. Although the techniques we discuss here can be applied broadly, our emphasis in the applications is generally medically oriented.

Why is statistics or biostatistics important? Consider an example. Most people believe that it is beneficial in terms of cardiovascular health to exercise regularly and to eat a low-fat diet. It is this underlying belief that often motivates us to exercise on a day when we otherwise might skip exercising or to feel guilty when we consume a high-fat meal or maybe two desserts. Should we be motivated by this belief or truly feel guilty? Maybe this idea of exercising and eating right is just a rumor that got out of control!

Why do we believe that exercising and eating healthy is associated with better cardiovascular health? Well, a number of research studies have shown that people who exercise regularly and eat low-fat diets are less likely to suffer from cardiovascular disease. An example of one such study is the Framingham Heart Study (Kannel, W. B.: "Habitual level of physical activity and risk of coronary heart disease: the Framingham Study."*Can. Med. Assoc. J.* 1967; 96: 811–812; Dawber, T. R., Kannel, W. B., Pearson, G., Shurtleff, D.: "Assessment of diet in the Framingham Study: methodology and preliminary observations. *Health News* 1961; 38: 4–6.). The Framingham Heart Study started in 1948 and involved over 5,000 men and women, who make up the Framingham cohort. At the beginning, each participant in the study had a complete physical examination and they came back for repeat examinations every 2 years. Along with monitoring such things as blood pressure, cholesterol, exercise, and nutrition, the study investigators also monitored whether participants had heart attacks or other cardiovascular disease over time. In the early 1970s, the offspring of the original cohort and their spouses were invited to join what is called the Framingham Offspring cohort. There are over 5,000 participants in the Offspring study and they have been examined approximately every 4 years. In 2002, a third generation was enrolled, which allows for more detailed investigations of genetic risk factors. This multigenerational study is still going on today, and much of what we understand about cardiovascular disease is derived from this very important study.

The details of investigations based on important studies often get reported in the newspaper or recapped on television newscasts. Should we believe that just because there was an association between exercise and cardiovascular disease among the participants in the Framingham Heart Study (or some other study) the same would hold for us? This idea of generalizing or inferring associations that are observed in a group of participants under study to the population at large is the crux of statistics.

Every day we are inundated with information. Information comes from newspapers, magazines, books, television newscasts, and the Internet. We need to distinguish important information from not-so-important information. We need to determine when the results of research studies apply to us and when they do not. It is unlikely that many of us will ever participate in a research study. However, we rely on well-conducted research studies to inform us of important associations, such as those between certain behaviors (e.g., exercise and diet) and better health. Statistics will provide us the tools to analyze and interpret studies.

Our primary goal in this book is to provide background for readers to apply and appropriately interpret statistical applications in the medical field. We have three specific aims:

- to provide an overview of *statistical vocabulary*
- to describe *statistical methodology and interpretation* and
- to introduce *statistical computing techniques*

Throughout the book, we present vocabulary associated with each application along with explicit definitions of terms and concepts. We provide readers with the tools to implement appropriate statistical techniques along with sufficient detail to understand each concept in a broad sense.

We cover a variety of methodological topics, from descriptive statistics and probability theory to statistical inference. The majority of the book is dedicated to methods for statistical inference, including one- and two-sample tests for means and proportions, analysis of variance techniques, and correlation and regression. In each topic area, we discuss the methodology, including assumptions and statistical formulas, along with appropriate interpretation of results. We introduce and discuss applications through real examples, most of which are taken from our work in applied biostatistics. We have purposely selected examples involving relatively few subjects to illustrate computations while minimizing the actual computation time. All of the techniques described can in practice be applied to larger problems.

Statisticians almost always use the computer and one or another statistical computing package to implement statistical techniques. There are a number of statistical computing packages available and most can be implemented on any computing platform (e.g., personal computer, Macintosh, UNIX, mainframe). In this book, we use SAS (SAS Institute Inc., SAS® User's Guide: Statistics, Version 8.02. Cary, NC: SAS Institute Inc., 1999–2001) to illustrate

statistical computing techniques. In each chapter, we present SAS programming code and output for each technique. Our objectives are to familiarize the reader with both the format and content of computer output and to provide appropriate interpretation of results. In the statistical computing sections, we present programming techniques for interested readers. An introduction to the statistical computing package SAS is contained in Appendix A, which includes an overview of the SAS system, techniques for creating and executing SAS programs, and examples of various techniques for entering data. Throughout the statistical computing sections, we include advanced statistical computing techniques, in the sections marked *, for more sophisticated readers. These sections can be omitted, as appropriate; such omission will not disrupt the reader's understanding of the material.

However, even though statistical techniques are almost always implemented using statistical computing packages, we feel strongly that readers must master techniques by hand before moving to computer applications. We therefore focus on explicit computations throughout this book. Wherever possible, we illustrate techniques by hand and then perform the corresponding analysis using SAS.

We also provide examples of statistical analysis on data collected from the Framingham Heart Study Offspring cohort. The data set contains real data but is appropriately masked to protect participant confidentiality. The details of the data set are described in Appendix C.