Preface

The purpose of this text is to enable biomedical researchers to use a number of advanced statistical methods that have proven valuable in medical research. The past thirty years have seen an explosive growth in the development of biostatistics. As with so many aspects of our world, this growth has been strongly influenced by the development of inexpensive, powerful computers and the sophisticated software that has been written to run them. This has allowed the development of computationally intensive methods that can effectively model complex biomedical data sets. It has also made it easy to explore these data sets, to discover how variables are interrelated and to select appropriate statistical models for analysis. Indeed, just as the microscope revealed new worlds to the eighteenth century, modern statistical software permits us to see interrelationships in large complex data sets that would have been missed in previous eras. Also, modern statistical software has made it vastly easier for investigators to perform their own statistical analyses. Although very sophisticated mathematics underlies modern statistics, it is not necessary to understand this mathematics to properly analyze your data with modern statistical software. What is necessary is to understand the assumptions required by each method, how to determine whether these assumptions are adequately met for your data, how to select the best model, and how to interpret the results of your analyses. The goal of this text is to allow investigators to effectively use some of the most valuable multivariate methods without requiring an understanding of more than high school algebra. Much mathematical detail is avoided by focusing on the use of a specific statistical software package.

This text grew out of my second semester course in biostatistics that I teach in our Masters of Public Health program at the Vanderbilt University Medical School. All of the students take introductory courses in biostatistics and epidemiology prior to mine. Although this text is self-contained, I strongly recommend that readers acquire good introductory texts in biostatistics and epidemiology as companions to this one. Many excellent texts are available on these topics. At Vanderbilt we are currently using Katz (2006) for biostatistics and Gordis (2004) for epidemiol-

ogy. The statistical software used in this text is Stata, version 10 (2007). It was chosen for the breadth and depth of its statistical methods, for its ease of use, excellent graphics and excellent documentation. There are several other excellent packages available on the market. However, the aim of this text is to teach biostatistics through a specific software package, and length restrictions make it impractical to use more than one package. If you have not yet invested a lot of time learning a different package, Stata is an excellent choice for you to consider. If you are already attached to a different package, you may still find it easier to learn Stata than to master or teach the material covered here from other textbooks. The topics covered in this text are linear regression, logistic regression, Poisson regression, survival analysis, and analysis of variance. Each topic is covered in two chapters: one introduces the topic with simple univariate examples and the other covers more complex multivariate models. The text makes extensive use of a number of real data sets. They all may be downloaded from my web site at `http://biostat.mc.vanderbilt.edu/dupontwd/wddtext/`. This site also contains complete log files of all analyses discussed in this text.

Changes in the Second Edition

I have made extensive modifications and additions to the second edition of this text. These can be summarized as follows:

- Since writing the first edition, Stata has undergone major improvements that make it much easier to use and enable more powerful graphics. The examples in this text take advantage of these improvements and comply with Stata's version 10 syntax.
- Stata now has easy-to-use point-and-click commands that may be used as an alternative to Stata's character-based commands. I have provided documentation for both the point-and-click and character-based versions of all commands discussed in this text.
- Appendix A summarizes the types of data discussed in this text and indicates which statistical methods are most appropriate for each type of data.
- Restricted cubic splines are used to analyze non-linear regression models. This is a simple but powerful approach that can be used to extend logistic and proportional hazards regression models as

well as linear regression models.
- Density-distribution sunflower plots are used for the exploratory analysis of dense bivariate data.
- The Breslow-Day-Tarone test is used to test the equality of odds ratios across multiple 2×2 tables
- Likelihood ratio tests of nested models are used extensively.
- I have added a brief discussion of proportional odds and polytomous logistic regression.
- Predicted survival and log-log plots are used to evaluate the adequacy of the proportional hazards model of survival data.
- Additional exercises have been added to several chapters.

Acknowledgement

I would like to thank Gordon R. Bernard, Jeffrey Brent, Norman E. Breslow, Graeme Eisenhofer, Cary P. Gross, Daniel Levy, Steven M. Greenberg, Fritz F. Parl, Paul Sorlie, Wayne A. Ray, and Alastair J. J. Wood for allowing me to use their data to illustrate the methods described in this text. I am grateful to William Gould and the employees of Stata Corporation for publishing their elegant and powerful statistical software and for providing excellent documentation. I would also like to thank the students in our Master of Public Health program who have taken my course. Their energy, intelligence and enthusiasm have greatly enhanced my enjoyment in preparing this material. Their criticisms and suggestions have profoundly influenced this work. I am grateful to David L. Page, my friend and colleague of 24 years, with whom I have learnt much about the art of teaching epidemiology and biostatistics to clinicians. My appreciation goes to Sarah K. Meredith for introducing me to Cambridge University Press; to William Schaffner and Frank E. Harrell, my chairmen during the writing of the first and second editions, respectively, who enabled my spending the time needed to complete this work; to W. Dale Plummer for programing and technical support with Stata and LaTeX; to Nicholas J. Cox for proof-reading this text, for his valuable advice, and for writing the *stripplot* program; to William R. Rising, Patrick G. Arbogast, and Gregory D. Ayers for proof-reading this book and for their valuable suggestions; to Jeffrey S. Pitblado for writing the Stata 8 version of the *sunflower* program and for allow-

ing me to adapt his LaTeX style files for use in this book; to Kristin MacDonald for writing the restricted cubic spline module of the *mkspline* program; to Tebeb Gebretsadik and Knut M. Wittkowski for their helpful suggestions; to ...?? ; and to my mother and sisters for their support during six critical months of this project. Finally, I am especially grateful to my wife, Susan, and sons, Thomas and Peter, for their love and support, and for their cheerful tolerance of the countless hours that I spent on this project.

W.D.D.
Nashville, Tennessee, U.S.A.

Disclaimer: The opinions expressed in this text are my own and do not necessarily reflect those of the authors acknowledged in this preface, their employers or funding institutions. This includes the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, USA.