Preface

Morris DeGroot died on November 2, 1989. His ideas did not die with him. I am honored to be allowed to carry on Morrie's tradition through the third edition of this text. Some changes have been made, but I have tried to preserve those features that make this book unique and attractive. Foremost among those features is Morrie's exceptional ability to explain things clearly. For this reason, the preface that you are about to read is as close to Morrie's own preface for the second edition as I could make it. I have changed those statements that are no longer true and explained the changes that were made to the third edition. But I wanted to give Morrie the opportunity to describe his book in his own words at least one more time.

This book contains more than enough material for a one-year course in probability and statistics. The mathematical requirements for the course are a knowledge of the elements of calculus and a familiarity with the concepts and elementary properties of vectors and matrices. No previous knowledge of probability or statistics is assumed.

The book has been written with both the student and the teacher in mind. Special care has been taken to make sure that the text can be read and understood with few obscure passages or other stumbling blocks. Theorems and proofs are presented where appropriate, and illustrative examples are given at almost every step of the way. More than 1200 exercises are included at the ends of the various sections of the book. Some of these exercises provide numerical applications of results presented in the text, and others are intended to stimulate further thought about these results. More than 250 supplementary exercises appear at the ends of the chapters.

The first five chapters are devoted to probability and can serve as the text for a onesemester course on that topic. The elementary concepts of probability are illustrated by such famous examples as the birthday problem, the tennis tournament problem, the matching problem, the collector's problem, and the game of craps. Standard material on random variables and probability distributions is highlighted by discussions of statistical swindles, the use of pseudo-random numbers, a comparison of the relative advantages of the mean and the median as predictors, the importance of the central limit theorem, and the correction for continuity. Also included as special features of these chapters are sections on Markov chains, the Gambler's Ruin problem, and utility and preferences among gambles. These topics are treated in a completely elementary fashion, but they can be omitted without loss of continuity if time is limited.

The next five chapters of the book are devoted to statistical inference. The coverage here is modern in outlook. Both classical and Bayesian statistical methods are developed in an integrated presentation. No single school of thought is treated in a dogmatic fashion. My goal is to equip the student with the theory and methodology that have proved to be useful in the past and promise to be useful in the future. These chapters contain a comprehensive but elementary survey of estimation, testing hypotheses, nonparametric methods, multiple regression, and the analysis of variance. The strengths and weaknesses and the advantages and disadvantages of such basic concepts as maximum likelihood estimation, Bayesian decision procedures, unbiased estimation, confidence intervals, and levels of significance are discussed from a contemporary viewpoint. Special features of these chapters include discussions of prior and posterior distributions, sufficient statistics, Fisher information, the delta method, the Bayesian analysis of samples from a normal distribution, unbiased tests, tests of goodness-of-fit, contingency tables, Simpson's paradox, robust estimation and trimmed means, confidence bands for a regression line, Bayesian analysis of simple linear regression, and residual analysis.

The main changes in this third edition are of eight types, which I summarize here.

- 1. A new Chapter 11 has been added on simulation. This includes methods for simulating specific distributions, importance sampling, Markov chain Monte Carlo, and the bootstrap. Special attention is paid to the importance of estimating the uncertainty in each simulation approximation. Examples are drawn from those problems that were difficult to solve analytically earlier in the text.
- 2. I have added a number of new examples. Many of these examples include actual data from published studies. The analyses of these data are *not* taken from the published sources, but rather are illustrations of the concepts covered in this text. Two good sources of data, including much that is described in this book, are the *Data and Story Library*, DASL, and Hand et al. (1994). The DASL Project is centered at Cornell University, and its data sets can be found on the worldwide web at http://lib.stat.cmu.edu/DASL/. The book by Hand et al. (1994) contains descriptions of about 500 data sets plus a diskette to save typing. Most of the examples from the second edition have survived to the third edition, but a few have been replaced. Some of the new examples illustrate the use of probability theory in genetics, queueing, and computational finance.
- 3. I have added some new sections or subsections on conditionally independent events and random variables, the lognormal distribution, quantiles, prediction and prediction intervals, improper priors, Bayes tests, power functions, *M*-estimators, residual plots in linear models, and Bayesian analysis of simple linear regression.
- 4. I have added a brief introduction and summary to each technical section. The introductory paragraphs are just to give readers a hint about what they are going to encounter. The summaries do not recap all of the material, but merely list the most important ideas. Some of the sections contain a lot of material, all of it useful, but some of it vital for moving on. Not all students are able to distinguish the vital from the merely useful on a first reading. (I hope that I have been able to distinguish them.) The summaries are designed to help highlight the vital topics.

- 5. I have added special notes at places where I felt it would be useful to make a brief summarization or a connection to a point made elsewhere in the text. These are indicated by the text NOTE: wherever they begin.
- 6. There has been some reorganization of material. Independence is now introduced after conditional probability. A couple of small sections from the second edition have be rewritten as notes or extended exercises. These include "Choosing the Best," "The Borel-Kolmogorv Paradox," "Inferences about the Median and Other Quantiles," and "The Regression Fallacy." All discussion of tables of random digits has been replaced by matrial on simulation. The two sections on sign and rank tests in the second edition have been combining into one section. The most serious reorganization occurs in Chapter 8. I have completely rewritten the first section introducing hypothesis testing. This section contains enough background material to allow immediate passage to a study of the t test in Section 8.5. The intervening sections on optimal tests are now optional. The material in these sections, although mathematically appealing, has lost some of its importance in modern treatments of statistics. I have expanded the coverage of power functions and introduced the noncentral t distribution. All discussion of Bayes tests now appears in a new Section 8.8.
- 7. Some material has been deleted from the second edition. In addition to those sections, mentioned earlier, that have had their main points summarized, I have removed the section on "Multidecision Problems." I also removed some discussions of other topics that I felt were not developed in sufficient depth to be useful at an introductory level. These include the likelihood principle, the Gauss-Markov theorem, and stepwise regression.
- 8. I have singled out some of the more challenging material from a number of sections and moved it to the ends of the sections. Special symbols indicate two categories of such material (I don't have the special symbols on my computer):
 - One symbol indicates the start of an extended example or in-depth discussion that is helpful in illustrating the wide applicability of the results. Some of this material can be a bit challenging on a first reading, however.
 - Another symbol indicates the start of a discussion of foundational issues that help develop a deeper understanding of the meanings of statistical techniques. This material is one of the unique features of the text, but it needs to be presented carefully.

If time does not permit complete coverage of the text, some of the following sections can be omitted without seriously affecting coverage of later sections: 2.4, 2.5, 4.9, 6.7, 6.8, 6.9, 7.6, 7.8, 8.2, 8.3, 8.4, 8.8, 8.9, 9.6, 9.7, 9.8, 10.4, 10.7, 10.8, and 11.4. Aside from cross-references between sections within this list, occassional material from elsewhere in the text does refer back to some of the sections in this list. Each of the dependencies is quite minor, however. Most of the dependencies involve references from Chapter 11 back to one of the optional sections. The reason for this is that the optional sections address some of the more difficult material, and simulation is most useful for solving those difficult problems that cannot be solved analytically. Except for passing references that help put material into context, I shall summarize the dependencies here.

- The sample distribution function (Section 9.6) is reintroduced during the discussion of the bootstrap in Section 11.5. The sample distribution function is also a useful tool for displaying simulation results. It could be introduced as early as Example 11.2.7 simply by covering the first subsection of Section 9.6.
- The material on robust estimation (Section 9.7) is revisited in some simulation exercises in Section 11.1 (Exercises 4, 5, 7, and 8).
- Example 11.2.4 makes reference to the material on two-way analysis of variance (Sections 10.7 and 10.8).

Although a computer can be a valuable adjunct in a course in probability and statistics such as this one, most of the exercises in the first ten chapters of this book do not require access to a computer or a knowledge of programming. (A few of the newly added exercises involving real data and residual plots are exceptions.) For this reason, the use of this book is not tied to a computer until one reaches the chapter on simulation. Even there, I have included a number of theoretical and methodological exercises for those who wish to study simulation as a statistical method akin to the ones described in the earlier chapters. Instructors are urged, however, to use computers in the course as much as is feasible. A small calculator is a helpful aid for solving some of the numerical exercises. Near the end of his life, Morrie began to show an interest in computers. This was in sharp contrast to when I first met him. When our department first installed computer terminals in all faculty offices, Morrie kept his under a dust cover. Later, when he saw that word processing software could produce text that looked as pleasing as that of a typewriter, he accepted the use of computers for word processing. In the years since Morrie's death, computers have begun to play a larger role in the development of statistical theory and methodology. I like to think that, had he lived, Morrie would have embraced the use of computers in teaching statistical theory and methods.

One further point about the style in which the book is written should be emphasized. The pronoun "he" was used throughout the second edition in reference to a person who is confronted with a statistical problem. Times have changed, and the editors have suggested alternative ways of referring to unspecified individuals. Throughout most of the text, the sex of a generic individual alternates between male and female from one example or exercise to the next. The field of statistics should certainly be accessible to both women and men as well as members of all minority and majority groups. It is my sincere hope that this book will help create among all groups an awareness and appreciation of probability and statistics as an interesting, lively, and important branch of science.

There are many people that I want to thank for their help and encouragement during this revision. First and foremost, I want to thank Marilyn DeGroot and Morrie's children for giving me the chance to revise Morrie's masterpiece. I am indebted to many readers, reviewers, colleagues, staff, and the people at Addison-Wesley whose help and comments have strengthened this edition. These include Paul Anagnostopoulos Brian Blank, Anthony Brockwell, Daniel Chambers, Cindy Cody, Michael Evans, Doug Frank, Lyn Geisler, Prem Goel, Joel Greenhouse, Susan Herring, Syed Kermani, Michael Lavine, John Lehoczky, John Liukkonen, Deirdre Lynch, Rosa Matzkin, Terry McConnell, Hans-Georg Muller, George Nichols, Mario Peruggia, Heidi Sestrich, Joe Snowden, Anna Stillner, HenSiong Tan, Kenneth Troske, Yehuda Vardi, Joseph Verducci, Mahbobeh Vezveai, Brani Vidakovic, Bette Warren, and Calvin Williams. If I left anyone out, it was unintentional, and I apologize. Errors inevitably arise in any project like this (meaning a project in which I am involved). For this reason, I shall post information about the book, including a list of corrections, on my web page http://www.stat.cmu.edu/ mark/ as soon as the book is published. Readers are encouraged to send me any errors that they discover.

The field of statistics has grown and changed since Morrie wrote a Preface for the second edition of this book in October, 1985. Personal computers (workstations) were just making the transition from expensive and curious toys to serious research tools. Popular statistical methods have been developed since 1985 which are heavily dependent on high-speed computation. Yet the length of time that a student can devote to the study of the field has not increased. Something has to give way. In recent years, there has been a trend toward teaching more of the computational methods and less of the mathematical theory. I won't pretend that this is a text on computational methods in statistics. But I have attempted to take a well-written text and modernize it enough to appeal to a new generation of statisticians. I hope that I have succeeded enough to justify the faith that Morrie's family has placed in me.

Pittsburgh, Pennsylvania June 2001 and October 1985 M.J.S. and M.H.D.