Preface

There is no doubt that the computer has revolutionized the practice of statistics in recent years. Computers allow us to analyze data more quickly using classical techniques, to analyze much larger data sets, to replace classical data analytic methods—whose assumptions may not be met—with more flexible computer intensive approaches, and to solve problems with no satisfactory classical solution.

Nor is there doubt that undergraduate mathematics and statistics courses could benefit from the integration of computer technology. Computer laboratories can be used to illustrate and reinforce important concepts; allow students to simulate experiments and visualize their results; and allow them to compare the results of classical methods of data analysis with those using alternative techniques. The problem is how best to introduce these techniques in the curriculum.

This book introduces an approach to incorporating technology in the mathematical statistics sequence, with an emphasis on simulation and computer intensive methods. The printed book is a concise introduction to the concepts of probability theory and mathematical statistics. The accompanying electronic materials are a series of in-class and take-home computer laboratory problems designed to reinforce the concepts and to apply the techniques in real and realistic settings.

The laboratory materials are written as *Mathematica* Version 5 notebooks [112] and are designed so that students with little or no experience in *Mathematica* will be able to complete the work. *Mathematica* notebooks contain text, data, computations, and graphics; they are particularly well suited for presenting concepts and problems and for writing solutions.

Laboratory problems, custom tools designed to enhance the capabilities of *Mathematica*, an introduction to using *Mathematica* for probability and statistics, and additional materials are included in an accompanying CD. An instructor's CD is available to those who adopt the book. The instructor's CD contains complete solutions to all laboratory problems, instructor guides, and hints on developing additional tools and laboratory problems.

The materials are written to be used in the mathematical statistics sequence given at most colleges and universities (two courses of four semester hours each or three courses of three semester hours each). Multivariable calculus and familiarity with the basics of set theory, vectors and matrices, and problem solving using a computer are assumed. The order of topics generally follows that of a standard sequence. Chapters 1 through 5 cover concepts in probability. Chapters 6 through 10 cover introductory mathematical statistics. Chapters 11 and 12 are on permutation

and bootstrap methods. In each case, problems are designed to expand on ideas from previous chapters so that instructors could choose to use some of the problems earlier in the course. Permutation and bootstrap methods also appear in the later chapters. Chapters 13, 14, and 15 are on multiple sample analysis, linear least squares, and analysis of contingency tables, respectively. References for specialized topics in Chapters 10 through 15 are given at the beginning of each chapter.

The materials can also be used profitably by statistical practitioners or consultants interested in a computer-based introduction to mathematical statistics, especially to computer intensive methods.

Laboratory problems

Each chapter has a main laboratory notebook, containing between five and seven problems, and a series of additional problem notebooks. The problems in the main laboratory notebook are for basic understanding and can be used for in-class work or assigned for homework. The additional problem notebooks reinforce and/or expand the ideas from the main laboratory notebook and are generally longer and more involved.

There are a total of 238 laboratory problems. Each main laboratory notebook and many of the problem notebooks contain examples for students to work before starting the assigned problems. One hundred twenty-three examples and problems use simulation, permutation, and bootstrap methods. One hundred twenty-five problems use real data.

Many problems are based on recent research reports or ongoing research—for example, analyses of the spread of an infectious disease in the cultured oyster population in the northeastern United States [18], [42], [100]; analyses of the ecological effects of the introduction of the Asian shore crab to the eastern United States [19], [20]; comparison of modeling strategies for occurrences of earthquakes in southern California [35]; comparison of spatial distributions of earthquakes [60] and of animal species [105]; comparison of treatments for multiple sclerosis [63], [8]; and analyses of associations between cellular telephone use and car accidents [88], between genetics and longevity [114], and between incidence of childhood leukemia and distance to a hazardous waste site [111]. Whimsical examples include comparisons of world-class sprinters [108] and of winning baseball players and teams [98].

Note to the student

Concepts from probability and statistics are used routinely in fields as diverse as actuarial science, ecology, economics, engineering, genetics, health sciences, marketing, and quality management. The ideas discussed in each chapter of the text will give you a basic understanding of the important concepts. The last section in each chapter outlines the laboratory problems.

Although formal proofs are not emphasized, the logical progression of the ideas in a proof is given whenever possible. Comments, including reminders about topics from calculus and pointers to where concepts will be applied, are enclosed in boxes throughout the text.

The accompanying CD contains two folders:

1. The `PDFFiles` folder contains documents in Acrobat PDF format. You will need a current copy of Adobe Acrobat Reader to open and print these files. Adobe Acrobat Reader is available for free from `adobe.com`.

2. The `MMAFiles` folder contains *Mathematica* files. You will need a copy of *Mathematica* Version 5 to work with these files.

The `PDFFiles` folder includes two appendices to the printed text and 15 laboratory workbooks. Appendix A is an introduction to the *Mathematica* commands used in the laboratory problems. Print Appendix A and keep it for reference. Appendix B contains tables of probabilities and quantiles suitable for solving problems when you are not using the computer. Print Appendix B and keep it for reference. There is one laboratory workbook for each chapter of the text. Print the ones you need for your course.

The `MMAFiles` folder includes 15 folders of laboratory problems and a folder of customized tools (`StatTools`). The `StatTools` folder should be placed in the user base directory or other appropriate directory on your system. Consult the online help within the *Mathematica* system for details, or speak to your instructor.

Note to the instructor

The material in the text is sufficient to support a problem-oriented mathematical statistics sequence, where the computer is used throughout the sequence. In fact, the first lab can be scheduled after three or four class meetings. Students are introduced to parametric, nonparametric, permutation, and bootstrap methods and will learn about data analysis, including diagnostic methods. (See the chapter outlines below.)

The text does not include exercises intended to be done by hand. You will need to supplement the text with by-hand exercises from other books or with ones that you design yourself. Suggestions for by-hand exercises that complement certain laboratory problems are given in the instructor's CD.

In addition, the printed text does not include *Mathematica* commands. Step-by-step instructions for using *Mathematica* commands are given in examples in the electronic materials. Online help is available, and Appendix A on the CD can be used as a reference.

Chapter outlines

Chapter 1 covers counting methods, axioms of probability, conditional probability, and independence. The first laboratory session is intended to be scheduled early in the term, as soon as the counting methods, axioms, and first examples are discussed. Students become familiar with using *Mathematica* commands to compute and graph binomial coefficients and hypergeometric probabilities (called "urn probabilities" in the lab) and get an informal introduction to maximum likelihood and likelihood ratio methods using custom tools. The additional problem notebooks reinforce these ideas

and include problems on frequency generating functions, conditional probability, and independence.

Chapters 2 and 3 are on discrete and continuous families of probability distributions, respectively. In the laboratory sessions, students become familiar with using *Mathematica* commands for computing probabilities and pseudorandom samples from univariate distributions, and with using custom tools for graphing models and samples. The additional problem notebooks reinforce these ideas, give students an informal introduction to goodness-of-fit, and include problems on probability generating functions, bivariate distributions, and transformations.

Chapter 4 is on mathematical expectation. In the laboratory and additional problem notebooks, students work with *Mathematica* commands for model and sample summaries, use sample summaries to estimate unknown parameters, apply the Chebyshev and Markov inequalities, and work with conditional expectations.

Chapter 5 is on limit theorems. In the laboratory session, students use custom tools to study sequences of running sums and averages, and answer a variety of questions on exact and approximate distributions of sums. The additional problem notebooks reinforce and expand on these ideas, and include several problems on probability and moment generating functions.

Chapter 6 serves as a transition from probability to statistics. The chi-square, Student t, and f ratio distributions are defined, and several applications are introduced, including the relationship of the chi-square distribution to the sampling distribution of the sample variance of a random sample from a normal distribution and the application of the chi-square distribution to the multinomial goodness-of-fit problem. In the laboratory session, students become familiar with chi-square and multinomial distributions, and use a custom tool for carrying out a goodness-of-fit analysis using Pearson's test (including analysis of standardized residuals). The additional problem notebooks contain simulation studies and applications of Pearson's goodness-of-fit test, and introduce students to minimum chi-square and method of moments estimates. The chapter is intended to precede formal statistical inference.

Chapters 7 and 8 are on estimation theory and hypothesis testing theory, respectively. In the first laboratory session, students become familiar with *Mathematica* commands for constructing confidence intervals for normal means and variances, and use custom tools to study the concepts of confidence interval and maximum likelihood estimation. In the second laboratory session, students become familiar with *Mathematica* commands for carrying out tests for normal means and variances, construct power curves, use a custom tool to construct tests and compute power at fixed alternatives, and compute sample sizes. The additional problem notebooks reinforce and expand on these ideas, contain simulation studies, introduce the idea of inverting tests to produce confidence intervals, and include applications of the likelihood ratio goodness-of-fit test.

Chapter 9 is on order statistics and quantiles. In the laboratory session, students apply custom tools for visualizing order-statistic distributions, for quantile estimation, and for constructing box plots in a variety of problems. The additional problem notebooks reinforce and expand on these ideas, introduce probability plots, study order statistics for uniform models, and contain simulation studies.

Chapter 10 is on parametric and nonparametric two sample analysis. In the laboratory session, students apply *Mathematica* commands for analyzing independent random samples from normal distributions and custom tools for the Wilcoxon rank sum test in a variety of problems. Normal probability plots of standardized observations are used to determine whether parametric methods should be used. The additional problem notebooks reinforce and expand on these ideas, contain simulation studies, introduce custom tools for quantile-quantile plots and inverting the Wilcoxon rank sum test under the shift model, and consider the randomization model for two sample analysis.

Chapter 11 is an introduction to permutation analysis, using nonparametric analyses of two samples and paired samples as first examples. In the laboratory session, students apply the rank sum, Smirnov, correlation, and signed rank tests in a variety of problems. The additional problem notebooks introduce a variety of different applications of permutation methods (using a variety of different test statistics) and use frequency generating functions to construct certain permutation distributions. Custom tools are used throughout, including tools for signed rank analyses, for constructing random reorderings of data, and for visualizing random reorderings of data.

Chapter 12 is an introduction to parametric and nonparametric bootstrap analysis. In the laboratory and additional problem notebooks, students consider the performance of the bootstrap and apply bootstrap estimation and testing methods in a variety of situations. Custom tools are used to construct random resamples, to visualize random resamples, to summarize the results of bootstrap analyses, and to construct approximate bootstrap confidence intervals using Efron's BC_a method in the nonparametric setting.

Chapter 13 is on parametric, nonparametric, and permutation methods for analysis of multiple samples. In the laboratory session, students use simulation to study analysis of variance for one-way layouts and blocked designs and to study Kruskal–Wallis and Friedman tests and apply these techniques in a variety of situations. Normal probability plots of standardized residuals are used to check analysis of variance assumptions. The additional problem notebooks reinforce these ideas and contain simulation studies and problems on analysis of variance in the balanced two-way layout setting. Custom tools are used throughout, including tools for analysis of variance, Bonferroni analysis, and Kruskal–Wallis and Friedman tests.

Chapter 14 is on linear least squares, including simple and multiple linear regression, permutation and bootstrap methods, and regression diagnostics. In the laboratory session, students use simulation to study the components of a linear regression analysis and apply the techniques in a variety of situations. The additional problem notebooks reinforce these ideas and contain problems on goodness-of-fit for simple linear models, analysis of covariance, model building, and locally weighted regression. Custom tools are provided for permutation analysis of slope in the simple linear setting, locally weighted regression, and diagnostic plots.

Chapter 15 is on large sample and small sample analyses of contingency tables, including diagnostic methods. In the laboratory session, students apply custom tools for large sample analyses of I-by-J tables and for constructing large sample confidence intervals for odds ratios to data from four studies. The additional problem

notebooks reinforce these ideas, consider the relationship between odds ratios and risk ratios, introduce McNemar's test for paired samples, and contain problems on permutation methods for fourfold and I-by-J tables.

Acknowledgments

This work was supported by Boston College through its faculty research programs and by the National Science Foundation through its Division of Undergraduate Education (NSF DUE 9555178). Boston College provided generous released time over several years while the materials were in development. NSF provided summer support for me, stipends for six additional faculty members, and generous support for an assistant.

Boston College Professors Dan Chambers and Charlie Landraitis, College of St. Catherine Professor Adele Rothan, C.S.J., and Stetson University Professor Erich Freedman used earlier versions of the laboratory materials in their classes and provided helpful comments and suggestions. Mt. Holyoke College Professor George Cobb and Harvard University Professor Marcello Pagano provided guidance on project design. I consulted with Boston College Professor Peg Kenney on assessment issues. University of Ottawa Professor John Nash and Boston College Professor Rob Gross provided interesting problem ideas and expert LaTeX advice. Ms. Sarah Quebec worked with me as an undergraduate and masters student at Boston College and then as a research assistant on this project; her thoughtful comments helped shape the final product. The comments provided by students in my classes were uniformly helpful in improving the laboratory materials. I extend a warm thank you to SIAM's editorial team, especially Linda Thiel, and to the reviewers of the text and laboratory materials.

Data sets were kindly provided by Fairfield University Biology Professor Diane Brousseau, Boston College Geophysics Professors John Ebel and Alan Kafka, Dr. Susan Ford (Haskin Shellfish Laboratory, Rutgers University), and Dr. Roxana Smolowitz (Marine Biological Laboratories, University of Pennsylvania).