Preface

In recent years developments in statistics have to a great extent gone hand in hand with developments in computing. Indeed, many of the recent advances in statistics have been dependent on advances in computer science and technology. Many of the currently interesting statistical methods are computationally intensive, either because they require very large numbers of numerical computations or because they depend on visualization of many projections of the data. The class of statistical methods characterized by computational intensity and the supporting theory for such methods constitute a discipline called "computational statistics". (Here, I am following Wegman, 1988, and distinguishing "computational statistics" from "statistical computing", which we take to mean "computational methods, including numerical analysis, for statisticians".)

The computationally-intensive methods of modern statistics rely heavily on the developments in statistical computing and numerical analysis generally.

Computational statistics shares two hallmarks with other "computational" sciences, such as computational physics, computational biology, and so on. One is a characteristic of the methodology: it is computationally intensive. The other is the nature of the tools of discovery. Tools of the scientific method have generally been logical deduction (theory) and observation (experimentation). The computer, used to explore large numbers of scenarios, constitutes a new type of tool. Use of the computer to simulate alternatives and to present the research worker with information about these alternatives is a characteristic of the computational sciences. In some ways this usage is akin to experimentation. The observations, however, are generated from an assumed model, and those simulated data are used to evaluate and study the model.

Advances in computing hardware and software have changed the nature of the daily work of statisticians. Data analysts and applied statisticians rely on computers for storage of data, for analysis of the data, and for production of reports describing the analysis. Mathematical statisticians (and even probabilists) use the computer for symbolic manipulations, for evaluation of expressions, for ad hoc simulations, and for production of research reports and papers. Some of the effects on statisticians have been subtle, such as the change from the use of "critical values" of test statistics to the use of "p-values", while others have been more fundamental, such as use of multivariate and/or nonlinear models instead of univariate linear models, which might formerly have been used as approximations because they were computationally tractable. More recently, computational inference using Monte Carlo methods are replacing asymptotic approximations. Another major effect that developments in computing have had on the practice of statistics is that many Bayesian methods that were formerly impractical have entered the mainstream of statistical applications.

The ease of computations has brought new attitudes to the statistician about the nature of statistical research. Experimentation has been put in the toolbox of the mathematical statistician. Ideas can be explored via "quick and dirty" computations. Ideas that appear promising after an initial evaluation can be pursued more rigorously.

Larger scale computing systems have also brought new attitudes to the statistician about the nature of discovery. Science has always moved ahead by finding something that was not being sought. Exploratory methods can be applied to very large datasets. Data mining of massive datasets has enabled statisticians to increase the rate of finding things that are not being looked for.

In computational statistics, computation is viewed as an instrument of discovery; the role of the computer is not just to store data, perform computations, and produce graphs and tables, but additionally to suggest to the scientist alternative models and theories. Many alternative graphical displays of a given dataset are usually integral features of computational statistics. Another characteristic of computational statistics is the computational intensity of the methods; even for datasets of medium size, high performance computers are required to perform the computations. Large-scale computations can replace asymptotic approximations in statistical inference.

This book describes techniques used in computational statistics, and considers some of the areas of application, such as density estimation and model building, in which computationally-intensive methods are useful. The book grew out of a semester course in "Computational Statistics" and various courses called "Topics in Computational Statistics" that I have offered at George Mason University over the past several years. Many of the various topics addressed could easily be (and are) subjects for full-length books. My intent in this book is to describe these methods in a general manner and to emphasize commonalities among them. The decomposition of a function so as to have a probability density as a factor is an example of a basic tool used in a variety of settings, including Monte Carlo (page 51), function estimation (Chapters 6 and 9), and projection pursuit (Chapters 10).

Most of the statistical methods discussed in this book are computationally intensive, and that is why we consider them to be in the field called computational statistics. As mentioned earlier, however, the attitude with which we embark on our analyses is a hallmark of computational statistics: the computations are often viewed as experiments and the computer is used as a tool of discovery.

I assume the reader has a background in mathematical statistics at roughly the level of an advanced undergraduate- or beginning graduate-level course in the subject, and, of course, the mathematical prerequisites for such a course, which include advanced calculus, some linear algebra, and the basic notions of optimization. Except for that prerequisite, the text is essentially self-contained.

Part I addresses in a general manner the methods and techniques of computational statistics. The first chapter reviews some basic notions of statistical inference, and some of the computational methods. The subject of a statistical analysis is viewed as a *data generating process*. The immediate object of the analysis is a set of data that arose from the process. A wealth of standard statistical tools are available for analyzing the dataset and for making inferences about the process. Important tools in computational statistics involve simulations of the data generating process. These simulations are used for *computational inference*. The standard principles of statistical inference are employed in computational inference. The difference is in the source of the data and how the data are treated.

The second chapter is about Monte Carlo simulation and some of its uses in computational inference, including Monte Carlo tests, in which artificial data are generated according to an hypothesis. Chapters 3 and 4 discuss computational inference using resampling and partitioning of a given dataset. In these methods a given dataset is used, but the Monte Carlo sampling is employed repeatedly on the data. These methods include randomization tests, jackknife techniques, and bootstrap methods, in which data are generated from the empirical distribution of a given sample, that is, the sample is resampled.

Looking at graphical displays of data is a very important part of the analysis of the data. The more complicated the structure of the data and the higher the dimension, the more ingenuity is required for visualization of the data; it is just in those situations that graphics is most important, however. Chapter 5 discusses methods of projecting data into lower dimensions, Chapter 6 covers some of the general issues in function estimation, and Chapter 7 presents a brief overview of some graphical methods, especially those concerned with multi-dimensional data. The orientation of the discussion on graphics is that of computational statistics; the emphasis is on discovery; and the important issues that should be considered in making presentation graphics are not addressed. The tools discussed in Chapter 5 will also be used for clustering and classification, and, in general, for exploring structure in data.

Identification of interesting features, or “structure”, in data is an important activity in computational statistics. In Part II, I consider the problem of identification of structure, and the general problem of estimation of probability densities. In simple cases, or as approximations in more realistic situations, structure may be described in terms of functional relationships among the variables in a dataset.

The most useful and complete description of a random data generating process is the associated probability density, and estimation of this special type of function is the topic of Chapter 8 and 9. If the data follow a parametric distribution, or rather, if we are willing to assume that the data follow a parametric distribution, identification of the probability density is accomplished by estimation of the parameters. Nonparametric density estimation is considered in Chapter 9.

Features of interest in data include clusters of observations and relationships among variables that allow a reduction in the dimension of the data. I discuss methods for identification of structure in Chapter 10, building on some of the basic measures introduced in Chapter 5.

Higher-dimensional data have some surprising and counterintuitive properties, and I discuss some of the interesting characteristics of higher dimensions.

Although statistical models are sometimes portrayed as overly limiting, and so methods thought to be "model-free" are sought, models play an important role throughout the field of statistics. In reality the model-free approach is just using a different, probably simpler, model. In Chapter 11, I discuss some of the standard models useful in statistics and consider the methods of building and analyzing models. Statistical modeling may be computationally intensive because of the number of possible forms considered or because of the recursive partitioning of the data used in selecting a model. In computational statistics, the emphasis is on *building* a model, rather than just estimating the parameters in the model. Parametric estimation of course plays an important role in building models.

As in Chapters 8 and 9, a simple model may be a probability distribution for some variable of interest. If, in addition, the relationship among variables is of interest, a model may contain a systematic component that expresses that relationship approximately and a random component that attempts to account for deviations from the relationship expressed by the systematic component.

A model is used by a statistician to assist in the analysis of data. In Chapter 11, I discuss the use of models in analyzing data, but I also often take the view that a model is a generation mechanism for data. A better understanding of a model can be assessed by taking this view; use the model to simulate artificial data, and examine the artificial data for conformity to our expectations or to some available real data. In the text and in the exercises of this chapter, I often use a model to generate data. The data are then analyzed using the model. This process, which is characteristic of computational statistics, helps to evaluate the *method* of the analysis. It helps us understand the role of the individual components of the model: its functional form, the parameters, and the nature of the stochastic component.

Monte Carlo methods are widely used in the research literature to evaluate properties of statistical methods. Appendix A addresses some of the considerations that apply to this kind of study. It is emphasized that the study uses an *experiment*, and the principles of scientific experimentation should be observed. Appendix B describes some of the software and programming issues that may be relevant for conducting a Monte Carlo study.

The exercises in this book contain an important part of the information that is to be conveyed. Many exercises require use of the computer, in some cases to perform routine calculations, and in other cases to conduct experiments on simulated data. The exercises range from the trivial or merely mechanical to the very challenging. I have not attempted to indicate which is which. Some of the Monte Carlo studies suggested in the exercises could be the bases for research publications.

When I teach this material, I use more examples, and more extensive examples, than what I have included in the text. Some of my examples form the basis for some of the exercises; but it is important to include discussion of them in the class lectures. Examples, datasets, and programs are available through links from the web page for this book.

In most classes I teach in computational statistics, I give Exercise A.3 in Appendix A (page 345) as a term project. It is to replicate and extend a Monte Carlo study reported in some recent journal article. Each student picks an article to use. The statistical methods studied in the article must be ones that the student understands, but that is the only requirement as to the area of statistics addressed in the article. I have varied the way the project is carried out, but it usually involves more than one student working together. A simple way is for each student to referee another student's first version (due midway through the term), and to provide a report for the student author to use in a revision. Each student is both an author and a referee. In another variation, I have students be co-authors. One student selects the article, designs and performs the Monte Carlo experiment, and another student writes the article, in which the main content is the description and analysis of the Monte Carlo experiment.

Software Systems

What software systems a person needs to use depends on the kinds of problems addressed and what systems are available. In this book I do not intend to teach any software system; and although I do not presume competence with any particular system, I will use examples from various systems, primarily S-Plus. Most of the code fragments will also work in R.

Some exercises suggest or require a specific software system. In some cases, the required software can be obtained from either `statlib` or `netlib` (see the Bibliography). The online help system should provide sufficient information about the software system required. As with most aspects of computer usage, a spirit of experimentation and of adventure makes the effort easier and more rewarding.

Software and "Reproducible Research"

Software has become an integral part of much of scientific research. It is not just the software system; it is the details of the program. A basic tenet of the scientific method is the requirement that research be reproducible by other scientists. The work of experimental scientists has long been characterized by meticulous notes describing all details that may possibly be relevant to the environment in which the results were obtained. That kind of care generally requires that computer programs with complete documentation be preserved. This requirement for reproducible research has been enunciated by Jon Claerbout (`http://sepwww.stanford.edu/`), and described and exemplified by Buckheit and Donoho (1995).

Taking care to preserve and document the devilish details of computer programs pays dividends not only in the communication with other scientists, but also for the person conducting the research. Most people begin writing programs before they become serious about their research; hence preservation and documentation are skills that must be acquired after bad habits have already developed.

A Word about Notation

I try to be very consistent in notation. Most of the notation is "standard". Appendix C contains a list of notation, but a general summary here may be useful. Terms that represent mathematical objects, such as variables, functions, and parameters, are generally printed in an italic font. The exceptions are the standard names of functions, operators, and mathematical constants, such as sin, log, E (the expectation operator), d (the differential operator), e (the base of the natural logarithm), and so on.

I tend to use Greek letters for parameters, and English letters for almost everything else, but in a few cases, I am not consistent in this distinction.

I do not distinguish vectors and scalars in the notation; thus, "x" may represent either a scalar or a vector, and x_i may represent the i^{th} element of an array, or it may represent the i^{th} vector in a set of vectors. I use uppercase letters for matrices, and the corresponding lowercase letters with subscripts for elements of the matrices.

I generally use uppercase letters for random variables and the corresponding lowercase letters for realizations of the random variables. Sometimes I am not completely consistent in this usage, especially in the case of random samples and statistics.

I thank the participants in our weekly seminar on computational statistics, in particular, Cliff Sutton, for enjoyable discussions on statistical methodology.

I thank John Kimmel of Springer for his encouragement and advice on this book and other books he has worked with me on. I also thank the reviewers for their comments and suggestions.

I thank my wife María, to whom this book is dedicated, for everything.

I used TeX via LaTeX to write the book, and I used S-Plus and R to generate the graphics. I did all of the typing, programming, etc., myself, so all mistakes are mine. I would appreciate receiving notice of errors as well as suggestions for improvement.

Material relating to courses I teach in the computational sciences is available over the World Wide Web at the URL,
`http://www.science.gmu.edu/`

Notes on this book, including errata, are available at
`http://www.science.gmu.edu/~jgentle/cmstbk/`

Fairfax County, Virginia

James E. Gentle
March 26, 2002

Springer

http://www.springer.com/978-0-387-95489-9

Elements of Computational Statistics
Gentle, J.E.
2002, XVIII, 420 p., Hardcover
ISBN: 978-0-387-95489-9