Preface

> Therefore reasoning does not suffice, but experience does.
>
> Mathematics is the door and key to the sciences.[1]
>
> —Roger Bacon (ca. 1214–1294)

Modern computers have revolutionized statistics. Techniques now routinely employed to analyze data were impractical and even unthinkable just a few years ago. Large data sets and exhaustive calculations can be handled comfortably, often by a typical laptop computer. Techniques once thought abstruse have become standard tools: principle component analysis, Markov chain Monte Carlo sampling, nonlinear model fitting, Bayesian statistics, Lomb-Scargle periodograms, and so on. Scientists and engineers must be conversant with more ways and more sophisticated ways to analyze data than ever before.

For many years I have taught a graduate course on data analysis for astronomers, physicists, and the occasional engineer. The purpose of the course has been to equip experimenters with the skills needed to interpret their data and theoreticians with enough knowledge to understand (and sometimes to question!) those interpretations. I was unable to find a book—or even a selection of books—that could serve as a textbook for the course. Much of the material in the course is not elementary and is not usually included in the many introductory books on data analysis. Books that did cover the material were highly specialized and written in a style and language opaque to most of my students. Books covering specific algorithms in specific computer languages were more appropriate as supplementary sources.

Driven by need, I wrote my own notes for the course, and the notes eventually grew to become this book. It is meant to be a book on advanced data analysis, not an introduction to statistics. Indeed, one may doubt whether yet another elementary introduction to, say, linear regression is really needed. At the same time, the book must be self-contained and must be understandable to readers with a wide variety of backgrounds, so it does cover basic concepts and tools. It includes many specific examples, but it is not a cookbook of statistical methods and contains not a line of computer code. Instead, the course and the

[1] "Ergo argumentum non sufficit, sed experientia." *Opus Maius*, Part 6, Chapter 1; "Et harum scientiarum porta et clavis est Mathematica." *Opus Maius*, Part 4, Chapter 1.

book emphasize the principles behind the various techniques, so that practitioners can apply the techniques to their own problems and develop new techniques when necessary. While the target audience is graduate students, the book should be accessible to advanced undergraduates and, of course, to working professionals.

The book is focused on the needs of people working in the physical sciences and engineering. Many of the statistical tools commonly used in other areas of research do not play a large role in the physical sciences and receive only minimal coverage. Thus, the book gives little space to hypothesis testing and omits ANOVA techniques altogether, even though these tools are widely used in the life sciences. In contrast, fits of models to data and analyses of sequences of data are common in the physical sciences, and Bayesian statistics is spreading ever more widely. These topics are covered more thoroughly.

Even so restricted, the subject matter must be heavily pruned to fit in a single book. The criterion I have used for inclusion is utility. The book covers data analysis tools commonly used by working physical scientists and engineers. It is divided into three main sections:

- The first consists of three chapters on probability: Chapter 1 covers basic concepts in probability, then Chapter 2 is on useful probability distributions, and finally Chapter 3 discusses random numbers and Monte Carlo methods, including Markov chain Monte Carlo sampling.
- The next section begins with Chapter 4 introducing basic concepts in statistics and then moves on to model fitting, first from a frequentist point of view (maximum likelihood, and linear and nonlinear χ^2 minimization; Chapters 5 and 6) and then from a Bayesian point of view (Chapter 7).
- The final section is devoted to sequences of data. After reviewing Fourier analysis (Chapter 8), it discusses power spectra and periodograms (Chapter 9), then convolution and image reconstruction, and ends with autocorrelation and cross correlation (Chapter 10).

An emphasis on error analysis pervades the book. This reflects my deep conviction that data analysis should yield not just a result but also an assessment of the reliability of the result. This might be a number plus a variance, but it could also be confidence limits, or, when dealing with likelihood functions or a Bayesian analysis, it could be plots of one- or two-dimensional marginal distributions.

Committed Bayesians may initially be unhappy that only one chapter is devoted to Bayesian statistics. In fact, though, the first two chapters on probability provide the necessary foundations for Bayesian statistics; and the third chapter, which includes a lengthy discussion of Markov chain Monte Carlo sampling, is motivated almost entirely by Bayesian statistics. Likelihood functions are covered extensively, albeit often tacitly, in the two chapters on least squares estimation. The book could easily serve as a textbook for a course devoted solely to Bayesian statistics. Because the book discusses both Bayesian and frequentist approaches to data analysis, it allows a direct comparison of the two. I have found that this comparison greatly improves students' comprehension of Bayesian statistics.

Nearly all the material in the book has already been published in other places by other people, but the presentation is my own. My goal has been to write about the material in a way that is understandable to my students and my colleagues. Much of the book is a translation from the elegant and precise language of mathematicians to the looser, workaday language of scientists and engineers who wrestle directly with data. The book nowhere mentions heteroskedastic data; it does discuss data with variable, sometimes correlated measurement errors!

The presentation is, nevertheless, mathematical, but the style is that of the physical sciences, not of mathematics. I have aimed for clarity and accuracy, not rigor, and the reader will find neither proofs nor lemmas. The book assumes that the reader is conversant with the calculus of several variables and is acquainted with complex numbers. It makes heavy use of linear algebra. It is my experience that most graduate students have taken at least one course on linear algebra, but their knowledge has rusted from lack of use, especially when it comes to eigenvalues and eigenvectors. A lengthy review of linear algebra is provided in appendix E. Specialized topics that would distract from the flow of the book have also been relegated to the appendices. Because of its crucial importance for the analysis of sequences, an entire chapter is devoted to Fourier analysis.

Finally, if you plan to learn or teach from this book, here is an observation, as valid today as 2400 years ago: "for the things we have to learn before we can do them, we learn by doing them, as builders by building and lyre players by playing the lyre."[2] To learn how to analyze data, analyze data—real data if available, artificial data if not. Pre-existing computer programs can readily be found for the analysis techniques discussed in the book, but, especially when first encountering a technique, use the programs only if you must and never without testing them extensively. It is better by far to write your own code.

Few people, certainly not I, can write a book like this without extensive help from colleagues, staff, and students. To all my students, current and former, especially to the students who have taken my course and given me feedback on earlier versions of the book; to my former postdocs, especially Allen Shafter, Janet (née) Wood, Coel Hellier, William Welsh, and Robert Hynes; and to my colleagues in the Department of Astronomy at the University of Texas, especially Terrence Deeming, William Jefferys, Pawan Kumar, Edward Nather, and Donald Winget, thank you.

[2] "ἃ γὰρ δεῖ μαθόνταζ ποιεῖν, ταῦτα ποιοῦντεζ μανθάνομεν, οῖον οικοδομοῦντεζ οικοδόμοι γίνονται καὶ κιθαρίζοντεζ κιθαρισταί." Aristotle, *Nicomachean Ethics*, Book II.