Preface

This book began as a collection of notes to myself, summarizing the details of various data analysis procedures that seemed like they would be useful in my job with the Central Research and Development Department of the DuPont Company. At the urging of the late W. David Smith, Jr., I began the task of turning those notes into this book and, more than any other single source, this book owes its completion to his persistent and extremely generous support. It is my overwhelming regret that he did not live to see the final manuscript.

Originally, this book was intended as a handbook for others with similar positions to mine, involved in the *exploratory analysis* of industrial process data, but as I became involved in a wider range of data analysis activities, the scope of the book broadened considerably. A more complete discussion of the term "exploratory analysis" is given at the beginning of Chapter 1, but the basic notion is the analysis of relatively unstructured datasets that arrive "without papers," accompanied by general questions, speculations, suppositions, or beliefs that *may or may not* turn out to be related to the available data. This type of analysis is to be contrasted with *confirmatory data analysis*, based on data obtained from *designed experiments*, that have been optimized specifically for the detailed questions at hand. My own experience and that of various others, like my former DuPont colleague Bob McClure, is that data analysis typically begins in an exploratory phase, often accompanied by a number of important misconceptions, and it is only after a period of exploratory analysis to sharpen the questions and uncover some of the more glaring of these misconceptions that we are ready for the traditional confirmatory analysis. Indeed, a careful exploratory analysis often provides both an excellent foundation for undertaking a confirmatory analysis, and sufficient motivation to do so.

As this book developed, redeveloped, and re-redeveloped more times than I care to admit, I expanded the scope to try to address the needs of others, particularly graduate students and academic researchers in engineering, the sciences and medicine. In this effort, I benefitted significantly from the opportunity to teach courses at the ETH (Swiss Federal Institute of Technology) in Zürich, based on early drafts of the book. My primary objective here is to provide a fairly comprehensive introduction to a range of broadly useful tools for exploratory data analysis, with a particular emphasis on the key working assumptions on which they are based. To this end, Chapter 1 begins with a broad overview of the topics covered in the rest of the book, along with some

indications of why these topics are practically important. Chapter 2 then gives a fairly broad overview of data: its types (e.g., real variables common to physics and engineering vs. the count and categorical variables common to the social sciences, medicine, and business), its description (i.e., metadata), and various important anomalies (e.g., outliers, inliers, missing values, duplicate records, and misalignment errors, among others). Next, Chapters 3 and 4 introduce discrete and continuous random variable models to describe uncertainty in these different data types, and Chapter 5 examines the problem of fitting straight lines to points in the plane, illustrating the range of solutions and results possible even in this apparently simple setting. Chapter 6 gives a brief introduction to estimation theory, including the Central Limit Theorem, possibly one of the most powerful, practically useful, and widely misunderstood results in statistics. Chapter 7 then examines the outlier problem in detail, emphasizing three points: first, that even a few outliers can badly distort the results obtained by most "classical" data analysis methods; second, that outliers do arise in practice, so it is not safe to ignore them; and third, that the term "outlier" is *not* synonymous with "gross measurement error," as is too often assumed. Chapter 8 introduces a number of useful graphical techniques for characterizing the variables in a dataset, including quantile-quantile plots, Poissonness plots, histograms, and kernel density estimators.

With the obvious exception of Chapter 5 on fitting lines to points in the plane, the first eight chapters of the book are primarily concerned with the characterization of individual variables, taken one at a time. Chapters 9 through 15 shift this focus to an exploration of relationships between variables, particularly those described by different types of regression models that predict the value or other characteristics of one variable from others. Specifically, Chapter 9 provides a brief introduction to the key ideas of hypothesis testing and confidence intervals, motivated by their importance in building and validating data models. Chapter 10 then considers the question of how to measure the degree of association between two variables (i.e., do they tend to vary in the same direction, in opposite directions, or independently?). Chapter 11 takes up linear regression models, generalizing the line-fitting problem considered in Chapter 5, and Chapter 12 examines the question of what happens when we apply transformations, examining data in a form that is different from the one in which we originally obtained it. As this chapter illustrates, variable transformations can have both very good and very bad effects. Chapter 13 extends the regression modeling ideas presented in Chapters 5 and 11 from the case where all variables are real-valued to the case of mixed data types, including the notions of odds ratios, ANOVA models, logistic regression models for binary responses, and Poisson regression models for count data. Chapter 14 introduces a number of broadly applicable ideas for characterizing data analysis results, including permutation methods, deletion diagnostics, and bootstrap resampling methods. Chapter 15 then applies these ideas to the regression models discussed in Chapters 11 and 13, presenting methods for assessing goodness-of-fit, deciding which variables to include in or exclude from a data model, detecting (possibly unreasonably) influential observations, and iterative model refinement.

Finally, Chapter 16 considers the important practical problem of missing data. Like outliers, missing data observations arise commonly in practice and their potential consequences are too important to be ignored. In favorable cases where only a few observations are missing from a large dataset, simple *imputation strategies* may be effective, where missing values are estimated from other, representative or related data observations. In unfavorable cases, however, more complex solutions are required, such as *multiple imputation* or techniques based on the *expectation-maximization algorithm*.

In my classes at the ETH, I covered most of the material relating to the analysis of real-valued data in a one semester graduate course for electrical engineering students (specifically, most of the material included in Chapters 1, 4, 5, 6, 7, 8, 10, 11 and 12). The basic prerequisite was a reasonable understanding of calculus and some familiarity with linear algebra. Most of the students also had some prior exposure to probability theory, but none was assumed and this book attempts to provide an approachable introduction to this material as it is needed.

One of my primary motivations in writing this book was that, as computing power becomes cheaper and data analysis software packages become easier to obtain and use, technical specialists in almost any area, from the anthropology of aardvarks to the zero-gravity x-ray crystallography of xenon compounds, can perform extensive computer-based analysis of observed data, *whether they understand the working assumptions on which those analyses are based or not, and whether they understand the consequences of violating those working assumptions or not.* The very real danger of this situation is that, most of the time, if we feed any particular dataset to any analysis procedure that does not reject it out of hand (e.g., due to gross data type mismatches), we will obtain numerical results. Hence, if we have set the problem up incorrectly, have included variables expressed in the wrong units, or have obtained a dataset that contains nonrepresentative values, we may obtain results whose uncritical acceptance can lead to disaster. A case in point is the 125 million dollar Mars Climate Orbiter, lost in 1999 because certain data values believed to be expressed in one set of units (newtons) were actually expressed in different units (pounds). To avoid such difficulties, it is necessary to critically examine both working assumptions and analytical results, and this book emphasizes both of these points.

Because most practical data analysis is done using commercially available or—increasingly—open-source software, a few words about software are in order. This book does not include detailed computational procedures, nor does it assume the reader will use any specific computational platform. The reason for this lack of specific computational focus—which may seem to some a glaring omission—is that the objective of this book is to help the reader set up and pursue a reasonable exploratory data analysis, *using whatever computational tools are available.* The examples presented in this book were prepared using a combination of the *S-Plus* commercial software package [272] and its open-source near-equivalent *R* [240]. As an aid to those wishing to work through the examples presented in this book on their own, instructions for obtaining *R* are presented in the Appendix at the end of the book.

For those wishing to use other platforms, it is important to emphasize that computational procedures very similar to those used here are widely available in other statistical software packages (e.g., *SAS*, *SPSS*, *STATA*, etc.), and that many of these procedures can easily be implmented in other environments like *Excel* or *MATLAB* that are not specifically statistical in nature, if they are not already available as built-in procedures. Conversely, it is also important to emphasize that even built-in procedures should be examined carefully before routine use. As an unfortunate example, Version 2.2 of the *MATLAB Statistics Toolbox* included a procedure called MAD that *appeared* to be the median absolute deviation from the median (MADM) scale estimate discussed in Chapter 7 and described by Huber [151, p. 107] as, "the single most useful ancillary estimate of scale." Unfortunately, what was actually implemented was the *mean absolute deviation from the mean,* a scale estimate with the same disasterous outlier sensitivity as the more familiar standard deviation [227]. Alternatively, since the median is available as a built-in procedure in the *MATLAB Statistics Toolbox*, construction of the MADM scale estimate based on the description given in Chapter 7 is a simple matter. The key point is that even software packages like *MATLAB* that are widely used and generally reliable can have some surprising features.

Although this book owes its greatest debt to Dave Smith, many others have also contributed significantly to its development and refinement. A complete list of acknowledgements would require another chapter, if not another volume. The incomplete list that follows must begin with Harmon Ray of the University of Wisconsin and my DuPont colleague Tunde Ogunnaike, both of whom supported my efforts with hours of discussion and detailed readings of preliminary draft chapters. In addition, Dave Smith's Advanced Process Control and Optimization Group had many academic visitors over the years, several of whom also contributed to the development of this book through their stimulating discussions on a wide variety of topics. Particularly significant contributions came from my interactions with Yaman Arkun, Frank Doyle, Dan Rivera, Jim Rawlings, and Dale Seborg. Similarly significant discussions with various DuPont colleagues should also be noted, including Bill Fellner, David Gay, Aaron Owens, Mike Piovoso, Martin Pottmann, and Dave Schnelle. I also owe a special acknowledgement to Bob McClure who introduced me to the practical aspects of analyzing real datasets under less than ideal circumstances. In addition, the final form of this book owes a great deal to the four extremely enjoyable years I spent at the Institut für Automatik at ETH, thanks to the director, Manfred Morari, and to Frank Allgöwer who initially invited me to ETH. Other ETH colleagues who improved this book in a variety of ways include Eric Bullinger, Cornelius Dorn, Rolf Findeisen, Andrea Gentillini, Thomas Güttinger, Daniel Lenz, Patrick Menold, Domenico Mignone, Boris Rankov, and Jan Ulrich. The inclusion of most of the material here on integer-valued and categorical data owes much to my subsequent involvement in the analysis of biomedical and clinical data, initially during a one-year position as Visiting Professor with the Tampere University of Technology in Finland as part of the Tampere International Center for Signal Processing, and subsequently with the Department of Pathol-

ogy, Anatomy and Cell Biology at Thomas Jefferson University in Philadelphia. Colleagues at those institutions to whom I owe a particular debt of gratitude include Jaakko Astola, Moncef Gabbouj and Olli Yli-Harja in Tampere, and Jim Schwaber and Greg Gonye at Thomas Jefferson University.

Finally, I offer my apologies to anyone I may have omitted from this list, and to the reader for any errors that may have worked their way into the final manuscript, despite my best efforts to exclude them.