Preface

My interest in missing data issues began in the early 1980s when I began working with the group that was to become the Institute for Health Promotion and Disease Prevention Research (better known as IPR) at the University of Southern California. This was my introduction to large-scale, longitudinal, field-experimental research. I had been trained in a traditional experimental social psychology program at the University of Southern California, and most of my colleagues at IPR (at least in the early days) happened also to have been trained as social psychologists. Given my training, much of my thinking in these early days stemmed from the idea that researchers had substantial control over the extraneous factors in their research. Thus, much of my early work was focused on gaining a degree of control in field experiment settings.

The challenges, of course, were numerous, but that is one of the things that made it all so interesting. One of the key challenges in those early days was missing data. The missing data challenge manifested itself as missing responses within a survey and as whole surveys being missing for some people at one or more waves of longitudinal measurement. Missingness within a survey often was due to problems with individual items (if students were confused by a question, a common reaction was to leave it blank) and problems with the length of the survey (slower readers would often leave one or more pages blank at the end of the survey). When whole surveys were missing from one or more waves of the longitudinal study, it was not uncommon that the student would return to complete a survey at a later wave. It was also common, however, that once a student was missing entirely from a measurement wave, that the student remained missing for the duration of the study.

In those days, there were no good analysis solutions for dealing with our missing data, at least none that one could expect to use with anything close to standard software. Our only real solution was to ignore (delete) cases with any missingness on the variables used for any statistical model. In fact, as I will discuss in Chap. 12, we even developed a planned missing data design (the first versions of the "3-form design") as a means of reducing the response load on our young student participants. Although this planned missing data design has proven to be an excellent tool for reducing response load, it further exacerbated our missing data analysis problems. My early

thinking on this was that because the pairwise-deletion correlations produced in this context would be random samples of the overall correlations, this would somehow help with our analysis problems. Although that thinking turned out to be correct, it wasn't for another 10 years that our analysis solutions would catch up.

I started thinking in earnest about missing data issues in the late 1980s. The impetus for this new thinking was that statisticians and other researchers finally began making good missing data analysis tools available. In fact, what happened in the missing data literature in 1987 alone can be thought of as a missing data revolution. In that single year, two major missing data books were published (Little and Rubin 1987; Rubin 1987). These two books were the statistical basis for most of the important missing data software developments in the following decade and beyond. Also published in 1987 were two influential articles describing a strategy for performing missing data analysis making use of readily available structural equation modeling (SEM) software (Allison 1987; Muthen et al. 1987). These articles were important because they described the first missing data analysis procedure that was truly accessible to researchers not trained as statisticians. Also published in 1987 was the article by Tanner and Wong (1987) on data augmentation, which has become a fundamental part of some approaches to multiple imputation.

Philosophy Underlying This Book

I feel it is important to give this brief history about the development of missing data theory and analysis solutions as well as the history of the development of my own skills in missing data analysis. It is important because my knowledge and experience in this area stemmed not from a background in statistics but from the need to solve the real problems we faced in the burgeoning discipline of prevention science on the 1980s and 1990s.

Because of my beginnings, my goals have always been to find practical solutions to real-world research problems. How can I do a better job of controlling the extraneous factors in a field experiment? How can I draw more valid conclusions about the success or failure of my intervention? Also, because I was trained as an experimental social psychologist and not as a statistician – not even as a quantitative psychologist – my understanding of the statistical underpinnings of various missing data techniques has often been couched in practical needs of the research, and my descriptions of these techniques and underpinnings have often relied more on plain English than on terms and language common in the statistical literature.

This practical basis for my understanding and descriptions of missing data techniques has caused some problems for me over the years. Occasionally, for example, my practical approach produces a kind of imprecision in how some of these important topics are discussed. To be honest, I have at times bumped heads a little with statisticians, and psychologists with more formal statistical training. Fortunately, these instances have been rare. Also, it has been my good fortune to have spent several years collaborating closely with Joe Schafer. This experience has been a huge benefit to my understanding of many of the important topics in this book.

On the other hand, my somewhat unusual, practical, approach to missing data techniques and underpinnings has gradually given me the ability to describe these things, in plain English, with a satisfying degree of precision. Further, my take on these issues, because it is so firmly rooted in practical applications, occasionally leads to discoveries that would not necessarily have been obvious to others who have taken a more formal, statistical, approach to these topics.

The long and short of this is that I can promise you, the reader, that the topics covered in this book will be (a) readable and accessible and (b) of practical value.

Prerequisites

Most of the techniques described in this book rely on multiple regression analyses in one form or another. Therefore, I assume that the reader will, at the very least, already have had a course in multiple regression. Even better would be that the reader would have had at least some real-world experience in using multiple regression. As I will point out in later chapters, one of the most flexible of the missing data procedures, multiple imputation, requires that the output of one's statistical analysis be a parameter estimate and the corresponding standard error. Multiple regression fits nicely into this requirement in that one always has a regression coefficient (parameter estimate) and a standard error. Other common procedures such as analysis of variance (ANOVA) can be used with multiple imputation, but only when the ANOVA model is recast as the equivalent multiple regression model.

Knowledge of SEM is not a prerequisite for reading this book. However, having at least a rudimentary knowledge of one of the common SEM programs will be very useful. For example, some of the planned missing data designs described in Section 4 of this book rely on SEM analysis. In addition, my colleagues and I have found the multiple-group SEM (MGSEM) procedure (Allison 1987; Muthen et al. 1987) to be very useful in the missing data context. The material covered in Chaps. 10 and 11 relies heavily on these techniques. Finally, knowledge of one of the major SEM packages opens up some important options for data analysis using the full information maximum likelihood (FIML) approach to handling missing data.

Because my take on handling missing data is so firmly rooted in the need to solve practical problems, or perhaps because my understanding of missing data theory and practice is more conceptual than statistical, I have often relied on somewhat low-tech tools in my solutions. Thus, I make the assumption that readers of this book will have a good understanding of a variety of low-tech tools. I assume that readers are well versed in the Microsoft Windows operating system for PCs.[1] For example, it will be extremely helpful if readers know the difference between ASCII

[1]I know very little about the operating system for Apple computers, but with a few important exceptions (e.g., that NORM currently is not available for Apple computers), I'll bet that good knowledge of the Apple operating system (or other operating systems, such as Unix or Linux) will work very well in making use of the suggestions described in this book.

(text) files (e.g., as handled by the Notepad editor in Windows) and binary files (e.g., as produced by MS Word, SAS, SPSS, and most other programs). Although the Notepad editor for editing ascii/text files will be useful to an extent, it will be even more useful to have a more full-featured ascii editor, such as UltraEdit (http://www.ultraedit.com).

Layout of this Book

In Section 1 of this book, Chaps. 1 and 2, I deal with what I often refer to as missing data theory. In Chap. 1, I lay out the heart of missing data theory, focusing mainly on dispelling some of the more common myths surrounding analysis with missing data and describing in some detail my take on the three "causes" of missingness, often referred to as missing data mechanisms. I also spend a good bit of space in Chap. 1 dealing with the more theoretical aspects of attrition. In Chap. 2, I describe various analysis techniques for dealing with missing data. I spend some time in this chapter describing older methods, but I stay mainly with procedures that, despite being "old," are still useful in some contexts. I spend most of the space in this chapter talking about the more theoretical aspects of the recommended methods (multiple imputation and maximum likelihood approaches) and the EM algorithm for covariance matrices.

In Section 2, I focus on the practice of multiple imputation and analysis with multiple imputed data sets. In Chap. 3, I describe in detail multiple imputation with Schafer's (1997) NORM 2.03 program. Chapter 4 covers analysis of NORM-imputed data sets with SPSS (versions 15, 16, and lower; and newer versions without the new MI module). In this chapter, I outline the use of my utility for automating the process of analysis with multiple imputed data sets, especially for multiple regression analysis. In Chap. 5, I describe multiple imputation with the recently released versions of SPSS (version 17–20) that include the MI module. In this chapter, I describe the process of performing multiple imputation with small data problems, staying within the SPSS environment, and performing automated analysis with regression and logistic regression. I also describe the limitations of this initial SPSS product (through version 20) and suggest the preferable alternative of doing MI with NORM 2.03 (along with my automation utility for reading NORM-imputed data into SPSS), but performing analysis and automation with the quite excellent automation features newly available in SPSS 17 and later versions. In Chap. 6, I cover the topic of imputation and analysis with cluster data (e.g., children within schools). I describe analysis of multilevel data with SPSS 17–20 Mixed module and also with HLM 6. I also describe a feature of my automation utility for analyzing NORM-imputed data with HLM 6–7. In Chap. 7, I discuss in detail multiple imputation with SAS PROC MI. In this chapter, I provide syntax for analysis with PROC REG, PROC LOGISTIC, and PROC MIXED and describe the combining of results with PROC MIANALYZE.

In Section 3, I focus on the practicalities of dealing with missing data, especially with multiple imputation, in the real world. In Chap. 8, I address the issue of spotting and troubleshooting problems with imputation. In Chap. 9 (with Lee Van Horn and Bonnie Taylor), I address the major practical concern of having too many variables in the imputation model. In Chap. 10, I cover the topic of doing simulation work with missing data. Given the popularity of simulations for answering many research questions, it is important to address issues that arise in the conduct of simulations relating to missing data. In addition to a brief description of the usual Monte Carlo approach to simulations, I also outline a more compact, non-Monte Carlo, approach that makes use of the multiple-group capabilities of SEM programs. In this section, I describe simulations based on MCAR missingness (this approach is at the heart of the material covered in Chap. 9), but I also extend this work in an important way to describe an approach to non-Monte Carlo simulations with MAR and MNAR missingness. In Chap. 11 (with Linda M. Collins), I cover the important area of including auxiliary variables in one's model. This chapter focuses mainly on addressing the problems associated with participant attrition. It touches on the value of auxiliary variables for bias reduction, but focuses on recovery of lost statistical power. The chapter covers practical strategies for including auxiliary variables in MI and FIML models. I outline an automation utility for determining the benefit of including auxiliary variables under a variety of circumstances.

Section 4 of the book describes the developing area of planned missing data designs. These designs allow researchers to make efficient use of limited resources, while allowing meaningful conclusions to be drawn. Chapter 12 describes the theory and practical issues relating to implementation of the 3-form design, a kind of matrix sampling design. Chapter 13 (with Allison Shevock; nee: Olchowski) describes a design we have called two method measurement. In this chapter, we present the theory and practical issues of implementing this SEM-based design.