

Introduction

In this chapter we give a broad introduction to the problem of missing data. We provide a perspective on the topic, reviewing the main developments of the last century (Section 1.1), in the process paying special attention to the setting of clinical studies (Section 1.2). We examine the move towards more principled approaches, more elaborate modelling strategies and, most recently, the important role of sensitivity analysis (Section 1.3). Finally, we map out the developments and material that make up rest of the book (Section 1.4). In the next chapter we introduce the key sets of data that will be used throughout the book to illustrate the analyses.

1.1 FROM IMBALANCE TO THE FIELD OF MISSING DATA RESEARCH

It is very common for sets of quantitative data to be incomplete, in the sense that not all planned observations are actually made. This is especially true when studies are conducted on human subjects. Examples abound in epidemiologic studies (Piantadosi 1997; Clayton and Hills 1993; Green *et al.* 1997; Friedman *et al.* 1998), in clinical trials (Kahn and Sempos 1989; Lilienfeld and Stolley 1994; Selvin 1996), and in the social sciences, especially in sample surveys, psychometry, and econometrics (Fowler 1988, Schafer *et al.* 1993; Rubin 1987; Rubin *et al.* 1995), to name but a few areas.

Our focus in this book is on intervention-based clinical studies. We mean this in an inclusive sense, however, implying that the methodology presented may be appropriate outside this setting, for example in the context of epidemiological studies as well as experimental and observational data in non-human life sciences, including agricultural, biological, and environmental research.

Early work on the problem of missing data, especially during the 1920s and 1930s, was largely confined to algorithmic and computational solutions to the induced lack of balance or deviations from the intended study design. See, for example, the reviews by Afifi and Elashoff (1966) and Hartley and Hocking (1971). In the last quarter of the twentieth century, general algorithms, such as the expectation–maximization (EM: Dempster *et al.* 1977), and data imputation and augmentation procedures (Rubin 1987; Tanner and Wong 1987), combined with powerful computing resources, largely provided a solution to this aspect of the problem.

Rubin (1976) provided a formal framework for the field of incomplete data by introducing the important taxonomy of missing data mechanisms, consisting of *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). An MCAR mechanism potentially depends on observed covariates, but not on observed or unobserved outcomes. An MAR mechanism depends on the observed outcomes and perhaps also on the covariates, but not further on unobserved measurements. Finally, when an MNAR mechanism is operating, missingness does depend on unobserved measurements, perhaps in addition to dependencies on covariates and/or on observed outcomes. During the same era, the *selection model*, *pattern-mixture model*, and *shared-parameter model* frameworks were established. These are depicted schematically in Figure 1.1. In a selection model, the joint distribution of the i th subject's outcomes, denoted \mathbf{Y}_i , and vector of missingness indicators, written \mathbf{R}_i , is factored as the marginal outcome distribution and the conditional distribution of \mathbf{R}_i given \mathbf{Y}_i . A pattern-mixture approach starts from the reverse factorization. In a shared-parameter model, a set of latent variables, latent classes, and/or random effects is assumed to drive both the \mathbf{Y}_i and \mathbf{R}_i processes. An important version of such a model further asserts that, conditional on the latent variables, \mathbf{Y}_i and \mathbf{R}_i exhibit no further dependence. Rubin (1976) contributed the concept of *ignorability*, stating that under precise conditions, the missing data mechanism can be ignored when interest lies in inferences about the measurement process. Combined with regularity conditions, ignorability applies to MCAR and MAR combined, when likelihood or Bayesian inference routes are chosen, but the stricter MCAR condition is required for frequentist inferences to be generally valid. A final distinction is made between missingness *patterns*. *Dropout* or *attrition* refers to the specific situation, arising in longitudinal studies, where subjects are observed without interruption from the beginning of the study until a given point in time, perhaps prior to the scheduled end of the study, when they drop out and do not return to the study. Given a rather strong focus in this book on longitudinal studies, dropout, an indicator of which is denoted by D_i , will occupy a prominent position. The general mechanism, where subjects can be observed and missing on any partition of the set of planned measurement occasions, is often called *non-monotone*

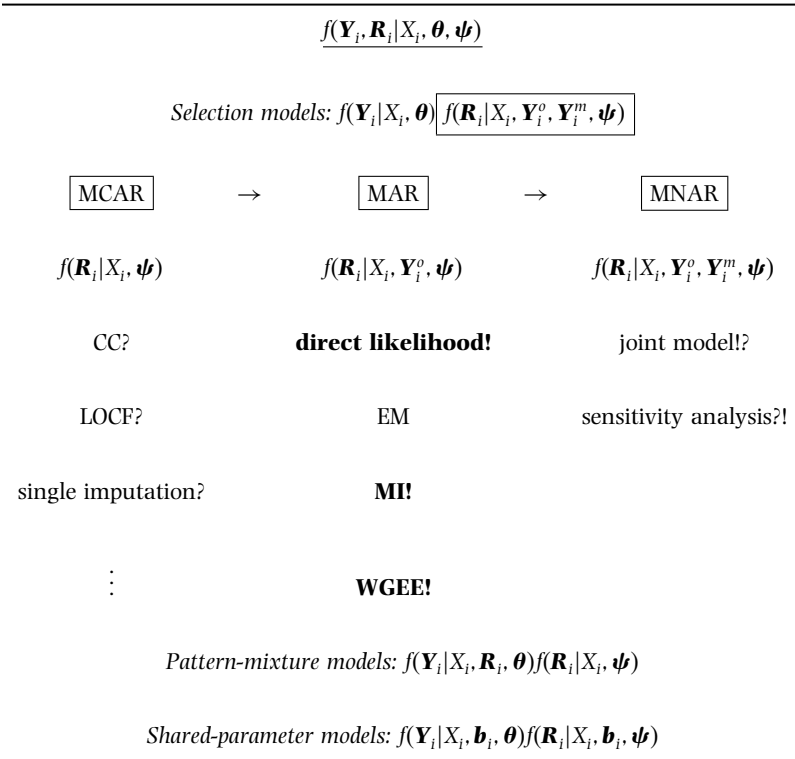


Figure 1.1 Schematic representation of the missing data frameworks and mechanisms, together with simple and more advanced methods, as well as sensitivity analysis. (MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random; CC, complete case analysis; LOCF, last observation carried forward; EM, expectation–maximization algorithm; MI, multiple imputation; WGEE, weighted generalized estimating equations.)

missingness. These and additional concepts are formalized and expanded upon in Chapter 3.

1.2 INCOMPLETE DATA IN CLINICAL STUDIES

In clinical trials, dropout is not only a common occurrence, there are also specific procedures for reporting and subsequently dealing with it. Patients who drop out of a clinical trial are usually listed on a separate withdrawal sheet of the case record form, with the reasons for withdrawal entered by the authorized investigator. Reasons frequently encountered are adverse events, illness not related to study medication, an uncooperative patient, protocol violation, and ineffective study medication. Further specifications may include so-called *loss*

to follow-up. Based on this medically inspired typology, Gould (1980) proposed specific methods to handle this type of incompleteness.

Even though the primary focus of such trials is often on a specific time of measurement, usually the last, the outcome of interest is recorded in a longitudinal fashion, and dropout is a common occurrence. While dropout, in contrast to non-monotone missingness, may simplify model formulation and manipulation, the causes behind it can be more problematic. For example, dropout may derive from lack of efficacy, or from potentially serious and possible treatment-related side effects. In contrast, an intermittently missing endpoint value may be due more plausibly to the patient skipping a visit for practical or administrative reasons, to measurement equipment failure, and so on. In addition, one often sees that incomplete sequences in clinical trials are, for the vast majority, of a dropout type, with a relatively minor fraction of incompletely observed patients producing non-monotone sequences. For all of these reasons we will put major emphasis on the problem of dropout, although not entirely neglecting non-monotone missingness in the process.

In a strict sense the conventional justification for the analysis of data from a randomized trial is removed when data are missing for reasons outside the control of the investigator. Before one can address this problem, however, it is necessary to establish clearly the purpose of the study (Heyting *et al.* 1992). If one is working within a *pragmatic* setting, the event of dropout, for example, may well be a legitimate component of the response. It may make little sense to ask what response the subject would have shown had they remained on study, and the investigator may then require a description of the response *conditional* on a subject remaining in the trial. This, together with the pattern of missingness encountered, may then be the appropriate and valid summary of the outcome. We might call this a conditional description. Shih and Quan (1997) argue that such a description will be of more relevance in many clinical trials. On the other hand, from a more *explanatory* perspective, one might be interested in the behaviour of the responses that occurred irrespective of whether we were able to record them or not. This might be termed a *marginal* description of the response. For a further discussion of intention to treat and explanatory analyses in the context of dropout see Heyting *et al.* (1992) and Little and Yau (1996), as well as Section 4.5 of this volume. It is commonly suggested (Shih and Quan 1997) that such a marginal representation is not meaningful when the nature of dropout (e.g., death) means that the response cannot subsequently exist, irrespective of whether it is measured. While such dropout may in any particular setting imply that a marginal model is not helpful, it does not imply that it necessarily has no meaning. Provided that the underlying model does not attach a probability of one to dropout for a particular patient, then non-dropout and subsequent observations are an outcome consistent with the model and logically no different from any other event in a probability model. Such distinctions, particularly with respect to the conditional analysis, are complicated by the inevitable mixture of causes behind missing values. The conditional description

is a mirror of what has been observed, and so its validity is less of an issue than its interpretation. In contrast, other methods of handling incompleteness make some correction or adjustment to what has been directly observed, and therefore address questions other than those corresponding to the conditional setting. In seeking to understand the validity of these analyses we need to compare their consequences with their aims.

Two simple, common approaches to analysis are (1) to discard subjects with incomplete sequences and (2) simple imputation. The first approach has the advantage of simplicity, although the wide availability of more sophisticated methods of analysis minimizes the significance of this. It is also an inefficient use of information. In a trivial sense it provides a description of the response conditional on a subject remaining in the trial. Whether this reflects a response of interest depends entirely on the mechanism(s) generating the missing values and the aims of the trial. It is not difficult to envisage situations where it can be very misleading, and examples of this exist in the literature (Kenward *et al.* 1994, Wang-Clow *et al.* 1995). Such imputation methods share the same drawbacks, although not all to the same degree. The data set that results will mimic a sample from the population of interest, itself determined by the aims of the analysis, only under particular and potentially unrealistic assumptions. Further, these assumptions depend critically on the missing value mechanism(s). For example, under certain dropout mechanisms the process of imputation may recover the actual marginal behaviour required while under other mechanisms it may be wildly misleading, and it is only under the simplest and most ignorable mechanisms that the relationship between imputation procedure and assumption is easily deduced. Little (1994a) gives two simple examples where the relationship is clear.

We therefore see that when there are missing values, simple methods of analysis do not necessarily imply simple, or even accessible, assumptions, and without understanding properly the assumptions being made in an analysis we are not in a position to judge its validity or value. It has been argued that while any particular *ad hoc* analysis may not represent the true picture behind the data, a collection of such analyses should provide a reasonable envelope within which the truth should lie. Even this claim is open to major criticisms, however, and we return to such ideas when sensitivity analyses are considered in Part III. In Chapter 4, after formally introducing of terminology and the necessary frameworks in Chapter 3, we provide a detailed examination of the advantages and drawbacks of simple methods, especially with a view to clinical trial practice.

As we explain in Chapter 4, it is unfortunate that so much emphasis has been given to methods such as *last observation carried forward* (LOCF), *complete case analysis* (CC), or simple forms of imputation. These are *ad hoc* methods defined *procedurally* in terms of manipulation of the data, rather than derived in a statistically principled way from the design of the trial and the aims of the analysis. As a consequence the relationship between their validity

and underlying assumptions can be far from clear and, when the relevant assumptions *can* be identified, they are seen to be very strong and unrealistic. In the LOCF procedure the missing measurements are replaced by the last one available. In particular, even the strong MCAR assumption does not suffice to guarantee that an LOCF analysis is valid. On the other hand, under MAR, valid inferences can be obtained through a likelihood-based or Bayesian analysis, without the need for modelling the dropout process. As a consequence, one can simply use, for example, linear or generalized linear mixed models (Verbeke and Molenberghs 2000; Molenberghs and Verbeke 2005), without additional complication or effort. This does not imply that these particular analyses are appropriate for all questions that might be asked of trial data, but the clarity of the underlying assumptions means that appropriate modifications can be readily identified when non-MAR analyses are called for, for example with intention to treat (ITT) analyses when dropout is associated with termination of treatment.

We will argue in Chapter 4, through the cases studies in Chapters 5 and 6, and then further throughout Part III, that such MAR-based likelihood analyses not only enjoy much wider validity than the simple methods but, moreover, are simple to conduct, *without additional data manipulation*, using such tools as the SAS procedures MIXED, GLIMMIX, or NLMIXED. Thus, clinical trial practice should shift away from the *ad hoc* methods and focus on likelihood-based ignorable primary analyses instead. As will be argued further, the cost involved in having to specify a model will arguably be mild to moderate in realistic clinical trial settings. Thus, we promote the use of direct likelihood ignorable methods and demote the use of the LOCF and CC approaches. Mallinckrodt *et al.* (2003a, 2003b), Molenberghs *et al.* (2004), and Lavori *et al.* (1995) propose direct likelihood and multiple imputation methods, respectively, to deal with incomplete longitudinal data. Siddiqui and Ali (1998) compare direct likelihood and LOCF methods.

1.3 MAR, MNAR, AND SENSITIVITY ANALYSIS

From the previous section, it is clear that not only is it advisable to avoid simple *ad hoc* methods such as complete case analysis and last observation carried forward, but there exists more appropriate flexible, broadly valid and widely implemented methodology. Principled methods and techniques such as direct likelihood and Bayesian analyses, the EM algorithm, multiple imputation, and weighted generalized estimating equations are systematically reviewed in Part IV. All of these methods are valid under the relatively relaxed assumption of MAR.

At the same time, it is important to consider reasons for departures from MAR, and the possible consequences of this for the conclusions reached. One obvious example, mentioned above, concerns treatment termination among dropouts in an ITT analysis. More generally, the reasons for, and implications

of, dropout are varied and it is therefore difficult, in fact usually impossible, to fully justify on a priori grounds the assumption of MAR. At first sight, this suggests a need for MNAR models. However, some careful considerations have to be made, the most important one of which is that no modelling approach, whether either MAR or MNAR, can recover the lack of information that occurs due to incompleteness of the data. These and related issues are given detailed treatment in Part IV.

An important feature of statistical modelling in the incomplete data setting is that the quality of the fit to the observed data need not reflect at all the appropriateness of the implied structure governing the unobserved data. This point is independent of the MNAR route taken, whether a parametric model of the type of Diggle and Kenward (1994) is chosen, or a semi-parametric approach such as in Robins *et al.* (1998). Hence in any incomplete data setting there cannot be anything that could be called a definitive analysis. Based on these considerations, it is advisable that, for primary analysis purposes, ignorable (MAR) likelihood-based methods be used. To explore the impact of deviations from the MAR assumption on the conclusions, one should then ideally conduct a sensitivity analysis, within which MNAR models can play a major role. The context of the trial and its aims can then be used to guide the choice of departures from MAR to explore in this sensitivity analysis.

The methods and strategies mentioned here have been added to Figure 1.1.

1.4 OUTLINE OF THE BOOK

In this book we place a strong emphasis on practical applications, and therefore we give more consideration to the explanation and illustration of concepts, and to proposing practicable modelling strategies, than to mathematical rigour. Excellent, rigorous accounts of missing data methodology can be found in Little and Rubin (1987, 2002), van der Laan and Robins (2003), and Tsiatis (2006). For these reasons, case studies feature prominently. In Chapter 2 we introduce a set of nine. The chapter also lists the places in the book where these sets of data are analysed so, if desired, their analyses can be followed in a step-by-step manner. In this chapter we also introduce eight additional examples which are used to further illustrate a range of points throughout the book.

In the final chapter of Part I we formalize the terminology, taxonomy, mechanisms, and frameworks introduced earlier in this chapter. This puts us in a position to provide a discussion of commonly used approaches in Part II and to provide a case for the need for a principled modelling approach (Chapter 4). The ideas laid down in this chapter are illustrated by means of two case studies, the orthodontic growth data (Chapter 5) and a series of depression trials (Chapter 6).

The MAR missingness mechanism, in particular in conjunction with ignorability, is the central idea in Part III. In Chapters 7, 8, and 9 we

deal with the direct likelihood approach, the EM algorithm, and multiple imputation. All three have a likelihood (or Bayesian) basis. The semi-parametric weighted generalized estimating equations (WGEE) technique is discussed in Chapter 10. In spite of their different background, it is appealing to combine multiple imputation and WGEE ideas (Chapter 11). A tangential but nevertheless interesting caveat in the direct likelihood method, originating from the frequentist nature of some inferences surrounding the likelihood, is the subject of Chapter 12. This part of the book concludes with the analysis of a case study in ophthalmology (Chapter 13) and an overview of how the methods dealt with in earlier chapters can be implemented using the SAS software system (Chapter 14).

In Part IV we explore modelling strategies available under MNAR, organized by modelling framework: selection models in Chapter 15, pattern-mixture models in Chapter 16, and shared-parameter models in Chapter 17. The specialized topic of protective estimation, referring to mechanisms where missingness depends on unobserved but not on observed outcomes, is studied in Chapter 18.

In Part V we turn to the very important subject of sensitivity analysis. This can roughly be divided into two. First, in Chapters 19 and 20, we use a variety of arguments to illustrate the nature of sensitivity when data are incompletely observed, and provide some arguments as to why this is the case. Second, a number of sensitivity analysis tools are presented. In Chapter 21 we are concerned with the so-called *region of ignorance* and *region of uncertainty* concepts, while in Chapter 22 we develop in some detail methods based on local and global influence. The nature and behaviour of local influence are further scrutinized in Chapter 23. In Chapter 24 we bring together selection, pattern-mixture, and shared-parameter model concepts in what is called a latent-class mixture model.

The concluding chapters that make up Part VI report on two case studies. In Chapter 25 the ophthalmology trial of Chapter 13 is revisited and subjected to a variety of analyses including assessment of sensitivity. In Chapter 26 we analyse quality-of-life data from a clinical trial with breast cancer patients.

We must emphasize that, in keeping with the theme of the book, our coverage should not in any sense be considered complete. The field of missing data is rapidly evolving, especially in the areas of more complex modelling strategies and sensitivity analysis. Therefore, we have settled for a wide range of instances of the various methods, selected because they make important points and/or can be used in practice. Throughout, we refer to key review papers, where the interested reader can often find more ample detail and further techniques not treated in this volume.