


RODERICK J. LITTLE AND DONALD B. RUBIN: STATISTICAL ANALYSIS WITH  
MISSING DATA

Wiley, Hoboken, NJ, 2020, 464 pp, \$98.95 (hardcover), \$79.00 (eBook), Print  
ISBN: 9780470526798, Online ISBN: 9781119482260

GERKO VINK 

UTRECHT UNIVERSITY

Few people have been as influential to the field of missing data analysis as Donald B. Rubin and Roderick J.A. Little. They have defined multiple generations in their way of thinking about incomplete data and the authors' message is approaching a widespread acceptance in contemporary data science. And although the first edition of their initiating book already appeared more than 30 years ago, the authors demonstrate that they are still able to influence the field. Once again, the authors produce the next iteration of their book which aims to give a non-exhaustive overview of the state of the art, complemented with novel methodology and new ways to think about the analysis of missing data.

The authors start right away by redefining the `obs` and `mis` notation that we have all grown so well-accustomed to. A subscript (0) now denotes the observed values and a subscript (1) denotes the missing values. Although they present valid arguments for this change, I have to admit that the change in notation did confuse me more than once when reading the book. That said, I do believe that it is this willingness to shake up the status quo that makes Little and Rubin's *Statistical Analysis with Missing Data (3rd edition)* still as relevant today as 3 decades ago.

**About the book:** *Statistical Analysis with Missing Data (3rd edition)* is a compilation of 15 chapters ranging over three parts: basic approaches, likelihood-based approaches and a part containing relevant practical examples of likelihood-based approaches to incomplete data. The authors concentrated on updating the almost 2 decades old 2nd edition, thereby focusing mostly on work they have been associated with. Even though this edition is not designed to cover everything, I have to admit that it still details an impressive array of approaches to handle a broad set of missing data problems. The order of the chapters naturally builds up in complexity and structure: when reading the book it feels like the reader is able to comprehend each chapter because it is preceded by the right precursor. The order of chapters and the topics therein remains the same as in the previous edition with the notable exception of the 15th chapter. This final chapter in the book has its focus shifted from non-ignorable missing data models to Missing Not at Random (MNAR) models. This different wording may seem trivial, but I am certain that this rewritten chapter would appeal to applied statisticians that have to deal with the often-feared missingness mechanism in practice. I am also delighted that the authors have extended the final chapter with discussions and demonstrations of techniques for sensitivity analyses and subsample regression. These topics and the corresponding examples will prove to be highly valuable in practice.

**Primary audience:** The third edition of Little and Rubin's *Statistical Analysis with Missing Data* is for the data analyst with both a passion for missing data analysis and a strong focus

Correspondence should be made to Gerko Vink, Department of Methodology and Statistics, Utrecht University, Padualaan 14, Utrecht, The Netherlands. Email: [G.Vink@uu.nl](mailto:G.Vink@uu.nl)

on statistics. Some chapters become quite technical and the nature of the discussion of topics requires a flexible attitude towards the sometimes inconsistent notation over the chapters. This book is therefore not an introduction for missing data analysis that would be ideal as a text for an undergraduate course. However, the book could serve well as the text for a graduate course. The problems formulated at the end of each chapters invite the reader to revisit the materials mostly by proving equations, deriving (posterior) distributions, writing likelihood functions or simulating data and missingness patterns. This means that this book is not for everyone. For example, applied researchers that would like to acquire the skills to solve their missing data problems in statistical software would benefit from a more practical book on incomplete data analysis. Examples of such practical books that would complement Little and Rubin's *Statistical Analysis with Missing Data (3rd edition)* are the books by Raghunathan et al. (2018) and Van Buuren (2018).

**Chapter overviews:** The first chapter sets up the nature of the problem. The authors begin by dependency between standard statistical techniques and rectangular data sets and proceed in Chapter 1 by introducing the nature of the problem. The authors detail the missingness mechanisms and patterns and take the reader by the hand through a rich set of examples of mechanisms that lead to missing data: such as censored data, missingness by design, attrition, measurement error and nonresponse. The authors conclude the first chapter with a taxonomy of missing data methods.

Chapter 2 details the first of basic approaches and focuses on missing data in experiments. The premise of this chapter is that the missing values create a misbalance in the carefully designed controlled experiments. The loss of balance means that techniques that require a complete outcome can no longer be used. The chapter details how to obtain correct least-squares estimates, standard errors, sums of squares and F tests when faced with missing data, starting with Bartlett's ANCOVA and gradually building up towards obtaining correct least-squares sums of squares with more than one degree of freedom.

Chapter 3 details complete case and available case analysis. A special appearance is reserved for weighting methods. Naturally, the chapter introduces some approaches to adjust for the bias that often comes when the analysis focuses on the observations alone.

Chapters 4 and 5 form a natural pair. Chapter 4 discusses single imputation methods are discussed aimed at obtaining point estimates of population quantities when data are incomplete. Imputation is the process of replacing the missingness with values and single imputation denotes that every missing datum is replaced by a single value. Chapter 5 is then aimed at solving the problem of too little variance due to filling in a single value for each missing datum. The authors detail estimates for uncertainty for the point estimates introduced in Chapter 4. Such uncertainty estimates are necessary as the additional variability due to the missingness should be taken into account in the variance of point estimates. Without adjustment, single imputed values would carry the same weight towards the final analyses as observed values would.

The second part of the book focuses on likelihood-based approaches. In Chapter 6 the authors implement inference based on the likelihood function on complete data and later on incomplete data. The authors distinguish between direct likelihood inference (amongst which Bayesian inference) and frequentist likelihood inference and consider each of these forms of inference in turn. The authors also define the partially Missing at Random (P-MAR) mechanism. The chapter ends with a small exploration of likelihood theory for coarsened data, i.e. data that are neither truly unobserved nor perfectly present in the complete-data sample space, such as, e.g. rounded or heaped numbers.

In Chapter 7, factored likelihood methods under ignorable missingness mechanisms are introduced. The chapter starts out with factored likelihoods for bivariate normal data where only one variable is subject to missingness. This scenario is carried forward from the previous chapter and it is nice to see that the authors recycle the same examples to illustrate new methods. There is a lot of room for discussing the application of drawing inference from data that have a monotone

missingness pattern and even some factored likelihoods to special non-monotone patterns are discussed. The authors also present a maximum likelihood and a Bayes computation for monotone normal data via the sweep operator.

Chapter 8 and 9 continue the likelihood-based focus of the second part in the book. Chapter 8 details the well-established expectation–maximization (EM) algorithm. The authors introduce the algorithm and detail the E step and the M step. The authors then gradually build up the chapter by discussing the properties of the algorithm and introducing extensions to the EM algorithm as well as hybrid maximization. The short Chapter 9 introduces the reader to some approaches to obtain inferences on large samples by means of maximum likelihood estimation.

Chapter 10 brings the reader to the widely applied Bayesian iterative imputation and multiple imputation methods. The chapter discusses the process of obtaining inferences from iterative imputation and details the need for a sufficient number of iterations in order to avoid that the simulations are not representative of the target distribution. In this edition, more focus is given to the popular chained equations approach to multiple imputation.

In the last part of the book, the authors give examples of the application of the introduced likelihood-based approaches. For example, in Chapter 11 Bayesian inference and multiple imputation techniques are applied to estimate the mean and covariance matrix from an incomplete multivariate normal sample, assuming that the missingness mechanism is ignorable. They also discuss a general repeated measures model that allows for missing data and discuss autoregressive models for univariate time series and Kalman filter models. At the end of Chapter 11, the authors revisit the relation between measurement error and missing data that was briefly introduced in Chapter 1. Chapter 12 focuses on robust modelling, i.e. modelling efforts that do not pose strict assumptions on the structure of the model or the form of the error distribution. Chapter 13 focuses on categorical data and Chapter 14 considers mixtures of normal and non-normal data under missing at random (MAR) assumptions.

Chapter 15 closes the book with a good discussion of methods that can be used when the missingness is assumed to be MNAR. Great detail is put into the discussion of specific MNAR models. A special focus is put on sensitivity analyses wherein the impact of deviations from the MAR assumption on the inference is studied. The authors have put great effort into rewriting this final chapter for the third edition of *Statistical Analysis with Missing Data* and have made the examples and discussion very relevant for contemporary data analysts.

**Overall conclusion** Overall, this is a well-organized book that provides comprehensive coverage of simple through advanced approaches for analysing missing data. The authors, who are perhaps the most established researchers in this field, provide us with valuable insight into the approaches of drawing inferences from incomplete data. That said, there are some reasons why this book would be a challenging introductory text. For example, the authors focus a lot on methods and procedures that can be difficult to apply in practice by non-expert statisticians and spend little time on iterative variable-by-variable imputation techniques that are becoming increasingly popular and are straightforward to apply. Despite this focus, I am confident that this book remains the de facto standard for researchers and graduate students that wish to develop a solid understanding of the statistical analysis of missing data.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### References

- Raghunathan, T., Berglund, P. A., & Solenberger, P. W. (2018). *Multiple imputation in practice: with examples using IVEware*. Boca Raton: CRC Press.
- Van Buuren, S. (2018). *Flexible imputation of missing data*. Boca Raton: CRC Press.

*Manuscript Received: 18 FEB 2022*  
*Final Version Received: 18 FEB 2022*  
*Published Online Date: 20 MAR 2022*