Statistical Science
2018, Vol. 33, No. 2, 139–141
https://doi.org/10.1214/18-STS332IN

Introduction to the Special Section on Missing Data

Julie Josse and Jerome P. Reiter

1. INTRODUCTION

Missing data is a problem for applied statisticians in every field. In survey-based inquiry, nonresponse rates have been increasing (National Research Council, 2013), threatening the validity of inferences from probability samples. In fact, some researchers argue that nonprobability samples and so-called found data, such as administrative databases from hospital or government files, may be preferable to probability samples riddled with missing values (Baker et al., 2013). Even found data, however, are not immune to missingness, as evidenced by reports of the impact of missing values in electronic medical records (e.g., Madden et al., 2016). Sometimes data are missing by design. For example, many analyses rely on databases constructed by fusing together multiple data sources, possibly with only a few observations in common. The combined data have large numbers of missing items.

There are numerous approaches to handling missing data (Little and Rubin, 2002). The most common approach, despite decades of research advocating otherwise, is to toss out the cases with missing values. At best, this is inefficient, as it wastes information from the partially observed cases. At worst, this can result in biased estimates, particularly when the distribution of the missing values is systematically different than the distribution of the observed values and rates of missingness are high. Fortunately, there are better alternatives to complete case analysis. Some analysts use model-based approaches, integrating likelihoods or posterior distributions over missing values. Some use imputation approaches, creating (multiple) completed datasets that can be subsequently analyzed. Some use weighting approaches, appealing to ideas from the design-based literature in survey sampling.

The aim of this special section of *Statistical Science* on missing data is to present a snapshot of some of the approaches to handling missing data, highlighting advances that have been made in recent years. It includes articles reviewing popular methodologies such as multiple imputation and double robust estimation. It also includes an article reviewing approaches when missing values are not ignorable. The section includes two articles connecting missing data to other areas of research, namely causal inference and low rank matrix completion, as both have strong ties to the missing data literature. The overarching aim is to promote the exchange of ideas from different perspectives on missing data.

Contributions come from leading researchers in missing data methodology and topical areas. We summarize each contribution in Section 2. The problems arising from missing values pervade most fields of application. As a consequence, the literature on missing data methodology is extremely rich. Naturally, one collection of articles cannot cover everything in missing data research. The topics covered here reflect our opinions on what we wanted to learn more about. We point to other topics in missing data research in Section 3.

2. SUMMARY OF ARTICLES

Multiple imputation is one of the most commonly used approaches to dealing with missing data. Murray's article, *Multiple Imputation: A Review of Practical and Theoretical Findings*, reviews several approaches for generating multiple imputations that use joint and conditional modeling, discussing pros and cons of each approach. He provides theoretical and empirical results in order to guide analysts in their choice of approach. Recent developments have focused on handling mixed type of variables, such as quantitative and categorical data, and on dealing with complex relationships between variables. Noticing that growing dimensionality demands growing complexity, Murray recommends Bayesian nonparametric mixture models to impute data. Such approaches have the advantage of naturally accounting for model uncertainty and ensuring proper imputation. Murray describes a truncated version of the Dirichlet process mixture of product multinomials for categorical data, and an approach

Julie Josse is Professor of Statistics, Ecole Polytechnique, route de Saclay, 91128 Palaiseau Cedex, France (e-mail: julie.josse@polytechnique.edu). Jerome P. Reiter is Professor of Statistical Science, Department of Statistical Science, Box 90251, Duke University, Durham, North Carolina 27705, USA (e-mail: jreiter@duke.edu).

based on two mixtures tied together further using a hierarchical structure for mixed data. He points to some extensions dealing with issues such as logically impossible cells and the presence of structural zeros.

Murray's article considers the setting of one dataset with observations that are independent and identically distributed. The article by Audigier and colleagues, *Multiple Imputation for Multilevel Data with Continuous and Binary Variables*, reviews multiple imputation methods for multilevel data. In particular, the authors describe approaches based on both joint modeling and fully conditional specifications using random effects regression models. They review proposals for handling both sporadically missing values, which correspond to some entries missing for some variables, and systematically missing values, for example, different subsets of variables are collected for different subsets of individuals. They use simulation studies to compare finite sample performances of these methods. The authors also provide guidance for practitioners on which methods to use for particular data settings.

Multiple imputation is not the only approach to handling missing data. The article by Seamans and Vansteelandt, *Introduction to Double Robust Methods for Incomplete Data*, reviews estimation approaches based on double robustnessness. These estimators have the appealing property that they are consistent under correct specification of either the model for missing data indicators or the model for the responses. Seamans and Vansteelandt describe ways to improve the efficiency of doubly robust estimators, for example by leveraging modern techniques such as the lasso and other regularization methods. They present connections to regression estimators of finite population quantities from design-based survey sampling.

Most missing data methods apply most naturally to data that are missing at random. Nonetheless, in many situations the data are missing not at random (MNAR). Analyses under MNAR require modeling the full-data distribution and the missingness mechanism, which typically is done using selection models, pattern mixture models, or shared parameter models. Linero's and Daniels's article, *A Bayesian Approach for Missing not at Random Outcome Data: The Role of Identifying Restrictions*, starts by briefly describing these approaches. Afterwards, the authors focus on identifying restrictions, which are crucial for many MNAR analyses as the distribution of the missing data given the observed data (called the extrapolation distribution) is not otherwise identifiable. Restrictions on the parameters depend on assumptions about the missing-data mechanism and are expressed in terms of conditional independence relationships. Linero and Daniels describe a broad range of identifying restrictions. They suggest an approach to deal with a MNAR outcome under a Bayesian framework. It combines a nonparametric Bayesian working model, which models only the distribution of the observed data, with identification restrictions. The Bayesian framework makes it easy to carry out sensitivity analyses.

The methods for missing values described in the first four articles—including weighting, imputation, doubly robust estimators, and sensitivity analysis—are also relevant for causal inference. This is particularly evident in the potential outcomes framework to causal inference (Rubin, 1974). In that framework, for each unit in the study one defines the outcome of interest under treatment and the outcome under control. The goal is to compare the two outcomes to learn the effect of the treatment; however, both outcomes can never be observed simultaneously, thus creating a missing data problem.

Ding's and Li's article, *Causal Inference: A Missing Data Perspective*, thoroughly reviews treatment effect estimation in the potential outcomes framework. They consider frequentist, Bayesian, and Fisherian paradigms, as well as discuss similarities and differences between missing data and causal inference perspectives. Their review of frequentist methods includes imputation, weighting, and doubly robust strategies. They highlight that building weights through covariate balance between two groups is extensively used in causal inference but not as much in missing value settings, whereas it is the other way around for doubly robust methods. The authors also discuss strategies to handle post-treatment variables that can be applied to settings with outcomes truncated by death. Their review of Bayesian methods includes a recommendation to use parametrizations like those described by Linero and Daniels, for example, separate parameters into those that can be estimated from the data and those that cannot be estimated from the data due to missingness. They end with a suggestion that theoretical developments in causal inference can be applied in statistical matching settings, that is, integrating datasets with two or more variables never observed jointly.

Imputation methods are at the root of matrix completion methods, which aim to recover as well as possible missing entries in large matrices. These methods are popular in machine learning communities and underlie many applications, such as recommender systems. Typically, matrix completion methods have not been embedded in a missing values framework, in the sense that they often assume missing completely at random data and almost never reflect the uncertainty in the imputed values. Clearly, connecting matrix completion methods and missing data paradigms is an area worthy of future research.

One exception to this characterization is Fithian's and Mazumder's article, *Flexible Low-Rank Statistical Modeling with Missing Data and Side Information*. They review methods and computational algorithms for modeling matrix-valued data using low rank assumptions, emphasizing matrix completion as an application. They describe common optimization routines and how they can break when analysts try to utilize side information about the rows or columns of the matrix. They present an alternative framework for modeling such matrices based on convex optimization with nuclear norm penalties. Fithian and Mazumder end their article with an explicit call for research on methods for MNAR data in matrix completion and related methods.

3. RESEARCH DIRECTIONS

Even though missing values research is intended toward applications, much research on missing data is done in relatively simple settings. For example, researchers assume only the outcome variable is missing and covariates are fully observed; they use illustrations with only a modest number of variables of the same type; use simple analyses models, or, they assume observations are independent and identically distributed. The theoretical validity of the results is often established only within a classical asymptotic framework (large n, fixed p). Some of the articles in this special section consider more complicated settings, but there is clearly a need to make stronger connections between theory and practice.

There are many important practical questions that require further developments. For example, what should be done with missing covariates if one wants to use doubly robust methods? How do we practically handle missing values that arise from different mechanisms, for example, some variables are MAR and others are MNAR, in a single coherent analysis? How should we do sensitivity analysis for MNAR data when using stochastic models with many parameters or optimization algorithms without obvious ways to separate identifiable and nonidentifiable parameters? How do we handle missing values in large-scale data with multimodal types—such as audio, geospatial, video, text, and traditional numeric data—for which traditional missing-data methodology were not necessarily designed?

New answers to these difficult questions may be found by investigating links across fields. For example, Mohan and Pearl (2018) are developing new approaches by casting missing data as a causal inference problem rather than the other way around. Techniques from machine learning, such as autoencoding and generative adversarial networks, may form the basis of new missing data methodology. Theoretical results, such as the work of Xie and Meng (2017) on uncongeniality in multiple imputation, may help practitioners use sound methods. New research on diagnostics and visualization may inform analyses with missing values.

Researchers in statistical science have developed many approaches for analyzing modern data. It is our sense that often these approaches have not been evaluated in the context of missing data. There is still a lot of work to be done, and we encourage others to work in this exciting field.

REFERENCES

BAKER, R., BRICK, J. M., BATES, N. A., BATTAGLIA, M., COUPER, M. P., DEVER, J. A., GILE, K. J. and TOURANGEAU, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* **1** 90–143.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken, NJ. MR1925014

MADDEN, J. M., LAKOMA, M. D., RUSINAK, D., LU, C. Y. and SOUMERAI, S. B. (2016). Missing clinical and behavioral health data in a large electronic health record (ehr) system. *Journal of the American Medical Informatics Association* **23** 1143–1149.

MOHAN, K. and PEARL, J. (2018). Graphical models for processing missing data. Technical Report, Dept. Computer Science, Univ. California, Los Angeles.

NATIONAL RESEARCH COUNCIL (2013). *Nonresponse in Social Science Surveys*. Panel on a research agenda for the future of social science data collection. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC.

RUBIN, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

XIE, X. and MENG, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? *Statist. Sinica* **27** 1485–1545. MR3701490