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

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).