International Journal of Behavioral Medicine (2020) 27:1–2
https://doi.org/10.1007/s12529-019-09834-y

INTEGRATIVE REVIEW

Statistical Guideline #4. Describe the Nature and Extent of Missing Data and Impute Where Possible and Prudent

Suzanne C. Segerstrom [1]

Published online: 6 December 2019

Abstract
From the Editors: This is one in a series of statistical guidelines designed to highlight common statistical considerations in behavioral medicine research. The goal is to briefly discuss appropriate ways to analyze and present data in the *International Journal of Behavioral Medicine* (*IJBM*). Collectively the series will culminate in a set of basic statistical guidelines to be adopted by *IJBM* and integrated into the journal's official Instructions for Authors, but also to serve as an independent resource. If you have ideas for a future topic, please email the Statistical Editor Suzanne Segerstrom at segerstrom@uky.edu.

Keywords Missing data · Imputation · Statistical guidelines

The Statistics Guru

Unless you are running a simulation study, you are likely to have missing data due to a skipped item or questionnaire page, a scale added after data collection has begun, a study dropout, or equipment failure, for example. The fourth statistical guideline for *IJBM* is a recommendation for authors to describe the nature and extent of their missing data and to impute missing data (that is, to replace missing data with a feasible value) where imputation is indicated.

The canonical question in missing data analysis is, what is the cause of missingness? Data can be missing completely at random (MCAR). For example, equipment might fail, causing a loss of heart rate data. A subset of questionnaires might have been copied incorrectly, leaving out a measure. Because the processes that generated the missing data had nothing to do with the nature of the research participants or their data, MCAR data do not risk biasing the results of analysis. Data can also be missing at random (MAR). For example, older participants might be more likely to drop out of a longitudinal study. In this case, the process that generated the missing data is related to a measured variable in the study. To reduce bias associated with MAR data, data analysis can account for the process by including the measured variable in the model. Data that are not missing at random (NMAR) are the most problematic and yield biased estimates. NMAR data are a function of the data that are missing (e.g., a person with a history of depression leaving questions about psychiatric history blank).

Many strategies for handling missing data exist, and both instructional articles [1–3] and book-length treatments are available; a good synopsis of books on missing data can be found at https://thestatsgeek.com/stats-books/missing-data-books/. This guideline cannot summarize all the approaches but suggests some reporting guidelines and possible starting points for handling missing data.

Missing Items It is not unusual for a person to skip an item or items in a questionnaire, and it is usual for investigators to take the mean of the remaining items rather than eliminate that person's data. This process, *ipsative mean imputation*, can work well if more items are present than are missing, the scale reliability is high ($\alpha > .70$), and the scale measures a single, well-defined domain [3]. Note that these guidelines apply to standardized items; if items are not standardized, then differences in mean levels among items can bias scores. Also note that this process only applies to item means; sums of items where there are missing data are biased because missingness will artificially deflate scores. (If the scale sum is desired, then the mean after ipsative imputation can be multiplied by the number of items.) This approach can be reported in a paper as

✉ Suzanne C. Segerstrom
segerstrom@uky.edu

[1] Department of Psychology, University of Kentucky, 125 Kastle Hall, Lexington, KY 40506-0044, USA

follows: "Ipsative mean imputation was used (N = [number of cases for which imputation was necessary]) when fewer than [criterion; should be < 50% [3]] of scale items were missing."

Missing Variables The amount of missingness for the study variables should be reported along with the data analysis description; for example, "variable *X* had 3% missing values and variable *Y* had 5% missing values, for a total of 7% of cases with any missing data and 1% of cases with both variables missing." This statement should be followed by a description of the mechanism of missingness, when known, and the strategy for addressing missingness. This strategy may be as simple as listwise deletion, in which cases with any missing data are not used in analysis. When data are MCAR and reasonable power is maintained, listwise deletion is a feasible option and does not yield biased estimates. When this is not the case, data analysis can be adjusted for bias (e.g., as in the MAR example above, age could be included in the models); data can be imputed via a number of different means; or, in longitudinal studies, models explicitly accounting for dropout can be implemented (e.g., pattern mixture models; [4]).

Imputation is a particularly useful strategy in behavioral medicine, in which the logistics of recruiting participants and collecting data can be more difficult and expensive than, for example, in questionnaire research with undergraduate students, and any data loss is costly. *Multiple imputation*, in which multiple datasets are imputed and then statistically combined, is a particularly useful and flexible strategy. Multiple imputation can make full use of a dataset, even when missing data rates are high (up to 50%) and N is small (N = 50) [2]. Contrary to what one hopes is a minority opinion, imputation is not "making up data"; it preserves the characteristics of the dataset and actually reduces bias relative to listwise deletion if data are not MCAR. Multiple imputation is a longstanding method that is available in many statistical packages. Rather than losing precious data, authors are encouraged to implement missing data procedures that will give more accurate results in their statistical models. No matter what authors choose to do, they should be transparent about their missing data, consider the consequences of different approaches, and handle missing data appropriately.

Compliance with Ethical Standards The author declares that she has no conflict of interest.

This article does not contain any studies with human participants or animals performed by the author, and so there was no requirement for informed consent.

References

1. Buhi ER, Goodson P, Beilands TB. Out of sight, not out of mind: strategies for handling missing data. Am J Health Behav. 2008;32:83–92.
2. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009;60:549–76.
3. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7:147–77.
4. Hedeker D, Gibbons RD. Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychol Methods. 1997;2:64–78.