On Missing Data*

What It Is, Why It's Important, and What You Should Do About It

Brooklin Becker

March 22, 2024

1 What Is Missing Data?

1.1 What Makes Data 'Missing'?

Missing data refers to variables which are missing or incomplete within research datasets (Osborne 2024). For a variable to be considering as missing data implies that the integrity of its measurement was not captured during initial data collection (Alexander 2023). As a result, the incompleteness or *missingness* of the variable preclude the intended inherent observation. Of course, some missing data may be *legitimate* wherein the absence of data is consequential of the observation being made (Osborne 2024). Conversely, *illegitimately* missing data occurs due to factors such as data collection or entry errors, technical or observational issues, or participant non-response, therefore interfering with the integrity of the measurement (Osborne 2024). We will focus on illegitimately missing data here forth.

1.2 Missingness

To best understand the relationship between successfully measured variables, and "the probability of missing data," (Baraldi and Enders 2010) *mechanisms of missingness* are integral to analysis. These mechanisms aim to explain why the data are missing, and how to best test its probability using missing data procedure (Osborne 2024). The mechanisms are as follows:

1. **Missing Completely at Random (MCAR)** - Data are MCAR when there is no systematic phenomenon which explains the missingness; it is simply happening at random (Little et al. 2014). It is rare for data to be MCAR, and it is certainly hard to empirically prove data are MCAR (Alexander 2023).

*Special thanks to Samantha Barfoot for her time and energy spent peer-editing this paper.

- For example, consider a situation where research is being done on the relationship between free menstrual products in the workplace and employee satisfaction. Due to time constraints and staffing shortages, some employees fail to complete or fill out the survey provided to them. Therefore, the reason for the missing data is unrelated to employee satisfaction based on availability of free menstrual products, but rather the employees' fiscal inability to complete the survey.

2. **Missing at Random (MAR)** - Data are MAR when there is a predictable or systematic phenomenon which explains its missingness (Little et al. 2014). As such, the impact of the missingness can be estimated and proven (Little et al. 2014).

- For example, suppose a shopping conglomerate is attempting to collect data on their top consumers. However, because the data collection involves questions detailing age, gender, and sexual orientation, some consumers abdicate from responding. In this situation, the missingness is not directly related to the unobserved values of the data but rather the comfort level of the individual responding as related to the observed variables.

3. **Missing Not at Random (MNAR)** - Data are MNAR when their missingness has a direct, systematic relationship to the value of that variable (Alexander 2023).

- For example, consider data collection being done on the amount of marijuana smokers in Salt Lake City, Ohio. Due to the prevalence of Mormonism in Salt Lake City, marijuana smokers may not disclose their use due to fear of religious consequences or social backlash. In this scenario, the usage of marijuana has a direct influence on whether or not the individual is willing to disclose the information.

2 Why is Missing Data Important?

Missing data, and the missingness *of* data, must be carefully considered whenever it is encountered. Missingness can be both meaningful, and indicative of a future analytical bias. Missingness can "[inform] and [reinforce] the status of a particular individual," or "produce answers that are biased, inefficient, and unreliable" (Osborne 2024). In other words, missing data can affect the integrity of the overall dataset and therefore any precision of conclusions made from it. The effect of the missing data must therefore be critically considered and analyzed in order to mitigate its impact (Osborne 2024).

3 What Should You Do About Missing Data?

Once mechanism of missingness is established, carefully consider which *missing data procedure* you plan to apply. For the purpose of this short paper, we will look at imputation methods as possible solutions to mitigate missing data, as they retain all initial data and reduce the

risk of false negatives in results (Bennett 2001). *Imputation methods* replace missing variables with possible values imputed by standardized methods (Bennett 2001).

3.1 Mean Imputation

Mean imputation is advantageous as a missing data procedure in a variety of cases as it is a relatively simple method. Mean imputation replaces missing variables with the mean of the observed or 'successful' value for that variable (Schafer and Graham 2002). To input the mean, a second data set must be constructed with the missing data removed so that the mean of such column can be calculated (Alexander 2023).

It is worth noting mean imputation does offer risk for bias wherein resulting estimates will often be either too large, or too small compared to the true value of the missing data (Little et al. 2014).

3.2 Multiple Imputation

By repeating the process of mean imputation (or other missing data procedures) several times, multiple imputation is achieved. As several different data sets are created and individually analyzed, uncertainty is quantified and can be accounted for in a statistical manner (Alexander 2023; Schafer and Graham 2002). Additionally, the pooling of these results from multiple imputations captures variability and yields more precise estimates of missing data parameters (Schafer and Graham 2002). Multiple imputation is therefore valuable in identifying "structural or casual relationships among variables" (Schafer and Graham 2002).

It is worth noting the validity of multiple imputation does rely on the integrity of the models and imputations created in the first place (Schafer and Graham 2002).

3.3 Your Role

Do take caution when dealing with missing data to recognize the role provenance plays in the ethical and responsible manipulation and analysis of data. No matter what missing data procedure you choose to employ, record every decision you make and every observation consequential to that decision to ensure transparency. Should you employ a missing data procedure, it is imperative that others are able to track your work from its infancy to its manipulated form in order to measure all factors which culminated such a result.

References

Alexander, Rohan. 2023. *Telling Stories with Data : With Applications in r*. First edition. Chapman & Hall/CRC Data Science Series. Boca Raton, FL: CRC Press, Taylor; Francis Group.

Baraldi, Amanda N., and Craig K. Enders. 2010. "An Introduction to Modern Missing Data Analyses." *Journal of School Psychology* 48 (1): 5–37. https://doi.org/10.1016/j.jsp.2009.10.001.

Bennett, Derrick A. 2001. "How Can i Deal with Missing Data in My Study?" *Australian and New Zealand Journal of Public Health* 25 (5): 464–69. https://doi.org/https://doi.org/10.1111/j.1467-842X.2001.tb00294.x.

Little, Todd D., Terrence D. Jorgensen, Kyle M. Lang, and E. Whitney G. Moore. 2014. "On the Joys of Missing Data." *Journal of Pediatric Psychology* 39 (2): 151–62. https://doi.org/10.1093/jpepsy/jst048.

Osborne, Jason. 2024. *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. Thousand Oaks: SAGE Publications, Inc. https://doi.org/10.4135/9781452269948.

Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2): 147–77. https://doi.org/10.1037/1082-989X.7.2.147.