What is Missing Data and What Should You Do About It?*

Chay Park

February 23, 2024

Missing data is a pervasive issue in research, data analysis, and statistical modeling that occurs when no data value is stored for a variable in an observation. It can significantly impact the validity, reliability, and generalizability of the conclusions drawn from a study. Understanding the nature of missing data, identifying its patterns and mechanisms, and adopting appropriate strategies to handle it are crucial steps in ensuring the integrity of research findings (Little & Rubin, 2002).

1 Types and Mechanisms of Missing Data

Missing data can be classified into three main types based on its mechanism: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR).

- **Missing Completely at Random (MCAR):** The probability of missingness is the same for all observations, implying that the missing data is unrelated to the observed or unobserved data. This is the least problematic type of missing data (Rubin, 1976).

- **Missing at Random (MAR):** The probability of an observation being missing is related to some of the observed data but not the missing data itself. In this case, the missingness can be accounted for by variables that have complete information (Schafer & Graham, 2002).

- **Missing Not at Random (MNAR):** The missingness is related to the value of the missing data itself. This type is the most challenging to handle because it requires making untestable assumptions about the missing data mechanism (Rubin, 1976).

*Code and data are available at: https://github.com/Chay-HyunminPark/Missing_data

2 Identifying Missing Data

The first step in addressing missing data is to conduct a thorough exploration of the dataset to identify the extent and pattern of missingness. This can involve visualizing missing data patterns, calculating the percentage of missing data for each variable, and assessing whether the data is MCAR, MAR, or MNAR. Understanding the nature of the missing data helps in selecting the most appropriate method for handling it (Van Buuren, 2007).

3 Strategies for Handling Missing Data

Several techniques are available for dealing with missing data, each with its advantages and limitations. The choice of method depends on the type and mechanism of missing data, as well as the research context.

- **Listwise Deletion (Complete Case Analysis):** This approach involves discarding any cases that have missing data on any of the variables of interest. While simple, it can lead to significant data loss and biased estimates if the data is not MCAR (Little & Rubin, 2002).

- **Pairwise Deletion:** Unlike listwise deletion, pairwise deletion uses all available data by analyzing pairs of variables that have no missing values. This method can retain more data but may lead to inconsistencies in the analysis (Schafer & Graham, 2002).

- **Imputation Methods:** Imputation involves replacing missing values with estimated ones. Simple imputation methods include using the mean, median, or mode of the observed data. More sophisticated techniques, such as multiple imputation, create several imputed datasets, analyze each dataset separately, and then combine the results. Multiple imputation can handle both MAR and MCAR data efficiently (Van Buuren, 2007).

- **Model-Based Methods:** These include using statistical models, such as maximum likelihood estimation (MLE) or Bayesian methods, to handle missing data. These approaches can provide unbiased estimates under MAR and MNAR conditions but require strong assumptions about the data (Schafer & Graham, 2002).

- **Sensitivity Analysis:** When dealing with MNAR data, conducting sensitivity analysis is crucial. This involves assessing how sensitive the results are to different assumptions about the mechanism of missingness (Little & Rubin, 2002).

4 Best Practices

The best approach to handling missing data is to prevent it as much as possible through careful study design and data collection strategies. However, when missing data is inevitable, it is essential to:

- Clearly document the extent and patterns of missingness.

- Choose the most appropriate method for handling missing data based on its mechanism.

- Conduct sensitivity analyses to assess the robustness of the findings.

- Report the methods used to handle missing data and their potential impact on the study conclusions (Donders et al., 2006).

In conclusion, missing data is a pervasive issue that requires careful consideration and strategic handling to ensure the credibility of research findings. By understanding the types and mechanisms of missing data and employing appropriate techniques to address it, researchers can mitigate its adverse effects and draw more reliable and valid conclusions from their analyses (Little & Rubin, 2002; Van Buuren, 2007).

References

Donders, A. R. T., G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons. 2006. "Review: A Gentle Introduction to Imputation of Missing Values." *Journal of Clinical Epidemiology* 59 (10): 1087–91.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Wiley.

Rubin, D. B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–92.

Schafer, J. L., and J. W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2): 147–77.

Van Buuren, S. 2007. "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification." *Statistical Methods in Medical Research* 16 (3): 219–42.