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

Hannah Yu

February 24, 2024

1 Understanding Missing Data and Strategies for Handling It

Missing data is a common issue encountered in data analysis across various fields, ranging from social sciences to healthcare and finance. It refers to the absence of values or information in a dataset for certain variables or observations. The presence of missing data can significantly impact the accuracy and reliability of statistical analyses and machine learning models. Therefore, it is essential for data analysts and researchers to understand the nature of missing data and employ appropriate strategies to handle it effectively.

2 Types of Missing Data:

Missing data can occur under various conditions with various reasons. It is crucial to have a good understanding of these reasons if one wishes to handle this situation strategically. Missing data can occur for various reasons, and understanding these reasons is crucial for determining the appropriate handling strategy. According to the classification of Vehtari, Gelman, and Hill (n.d.), known-missing observations, or observations we are aware that are missing, can be categorized into the following three type: Missing Completely At Random (MCAR), Missing at Random (MAR), and Missing Not At Random (MNAR).

Missing Completely At Random (MCAR): When data are missing completely at random, the observations that are missing are unrelated to any other variables regardless if the variable is in the dataset or not. Under this condition, the probability of data being missing is the same for all observations. For example, if survey responses are lost due to a technical error in data collection, it can be considered MCAR because the data lost is completely random.

*Code and data are available at: https://github.com/hannahyu07/what-is-missing-data

Missing at Random (MAR): Missing at Random (MAR) occurs when the probability of data being missing depends on other variables in the dataset but not on the missing data itself. In other words, the missingness is related to the observed data but not to the missing values. For example, if we are studying the effect of income and gender on political participation, after having gathered all three of the variable of interests, we noticed the pattern that male are less disclose about their income Alexander (2023).

Missing Not At Random (MNAR): Missing Not At Random (MNAR) happens when the probability of data being missing is related to the missing values themselves or to unobserved variables. In MNAR, the missingness is systematically related to the missing values, which can bias the analysis if not handled properly. For example, if respondents with higher income are less likely to disclose their income, it is MNAR because the missing values in income is related to income itself Alexander (2023).

3 Strategies for Handling Missing Data:

Dealing with missing data requires careful consideration and appropriate strategies to mitigate potential biases and inaccuracies in the analysis. We will elaborate on some of the common strategies for handling missing data.

Complete Case Analysis (CCA) is the most straightforward approach in handling missing data. CCA, also known as list-wise deletion, involves excluding observations with missing values from the analysis Ross, Breskin, and Westreich (2020). This approach is convenient, but it can lead to loss of valuable information and potential bias, especially if the missing data is not completely random.

Another common method to handle missing data is the imputation methods. This methods involve replacing missing values with estimated values based on the available data. One of the imputation methods is mean imputation, where researchers replace missing values by the mean of the observed values for that variable Glas (2010). The drawbacks of mean imputation is that it can underestimate the variability and lead to biased estimates, especially if the data are not MCAR. There are a variety of imputation methods including median imputation, mode imputation, regression imputation, and multiple imputation.

Multiple imputation is a branch of the imputation methods that are more sophisticated than mean imputation. "Multiple imputation fills in missing values by generating plausible numbers derived from distributions of and relationships among observed variables in the data set. Li, Stuart, and Allison (2015)" This method accounts for the uncertainty associated with imputed values and provides more accurate estimates compared to single imputation methods.

One of the most sophisticated and flexible method is the model-based methods. It requires involve using probabilistic models to predict missing values based on the observed data "What Are the Most Effective Methods for Imputing Missing Data in ML Models?" (n.d.). Some common model used in this method include linear regression, logistic regression, and machine learning algorithms. These methods can provide more accurate estimates, especially when the mechanism of the missing data is complex. However these methods are generally harder to interpret and requires more underlying assumptions and parameters.

4 Conclusion:

Missing data is a common challenge in data analysis, and addressing it effectively is essential for producing accurate and reliable results. In order to select appropriate handling strategies, it is important to understand the types of missing data and their implications.

References

Alexander, Rohan. 2023. "Telling Stories with Data." *Tellingstorieswithdata.com*. https://tellingstorieswithdata.com/.

Glas, C. A. W. 2010. "Imputation Method - an Overview | ScienceDirect Topics." *Www.sciencedirect.com*. https://www.sciencedirect.com/topics/mathematics/imputation-method#:~:text=Imputation%20methods%20are%20those%20where.

Li, Peng, Elizabeth A. Stuart, and David B. Allison. 2015. "Multiple Imputation." *JAMA* 314 (18): 1966. https://doi.org/https://doi.org/10.1001/jama.2015.15281.

Ross, Rachael K, Alexander Breskin, and Daniel Westreich. 2020. "When Is a Complete-Case Approach to Missing Data Valid? The Importance of Effect-Measure Modification." *American Journal of Epidemiology* 189 (12): 1583–89. https://doi.org/https://doi.org/10.1093/aje/kwaa124.

Vehtari, Andrew, Jennifer Gelman, and Aki Hill. n.d. "Regression and Other Stories." *Avehtari.github.io*. https://avehtari.github.io/ROS-Examples/.

"What Are the Most Effective Methods for Imputing Missing Data in ML Models?" n.d. *Www.linkedin.com*. https://www.linkedin.com/advice/0/what-most-effective-methods-imputing-missing-data.