Missing Data and Strategies to Manage Them*

Yuchao Niu (Reviewed by Ahnaf Alam)

March 5, 2024

Introduction

Missing data in statistics refers to the absence of values for one or more variables in a dataset, which is a common occurrence in social and medical research. Despite the effectiveness of our data acquisition processes, instances of missing data are likely to occur (Alexander 2023). A missing value, if observed, would hold significance for analysis; thus, a missing value can be viewed as a case where meaningful values are obscured (Little and Rubin 2019) Missing data can arise from a variety of causes, including data entry errors, equipment malfunctions, and participant non-response. Understanding the mechanisms behind missing data is crucial. Missing data can complicate the analysis due to the loss of information and the potential bias it introduces. Recognizing the nature of missing data and the strategies to manage it is vital to ensure the validity of study results. This paper discusses the various types of missing data and explores current strategies to mitigate the effects of missing data from the perspective of data management.

Types of Missing Data

Missing data are categorized by Donald B. Rubin into the following types: missing completely at random, missing at random and missing not at random (Rubin 1976).

Missing Complete At Random (MCAR) is when the absence of the data point occurs independently of any other variables in or outside of the dataset. One example of missing completely at random is when a survey participant unintentionally missed a question on the survey form. And the question that is missed has no relation with any other factors within and outside of the dataset of the survey. The absence of the data therefore will not introduce bias related to the missing data itself. Missing Complete At Random is quite rare to happen.

*Code and data are available at: https://github.com/MelanieNiu/Mini-essay-8

Missing At Random is when the probability of data being missing is related to some variables in the dataset but not related to the values of the missing data themselves. For example, in a survey where participants' income level and exercise level are investigated. Suppose the participants within a certain age group are less likely to report their income, data about the income level will be missing. However, the missingness is not related to the income itself but related to another observable variable age.

Lastly Missing Not At Random (MNAR) is when the probability of data being missing is related to an unobserved variable or the missing variable itself. There is an underlying mechanism that dictates whether the data will go missing. In a hypothetical scenario, respondents of a political survey are asked for the support of the Democratic Party, the respondents' education level is not collected, however respondents who has higher education level are less likely to respond. Another example is from the Youth Cohort Time Series data collected by the UK government. Students are shown less likely to report parental occupations when they fall into the category of 'intermediate level' or 'working level' compared to 'managerial level'. The missingness of the occupation data is related to the occupation itself. MNAR poses significant challenges in statistical analysis because it can introduce significant bias and makes it difficult to accurately estimate the parameter.

Strategies to Handle Missing Data

Addressing the issue of missing data is a necessary step in all statistical analyses and should be considered for different scenarios and types of data. Generally missing data can be addressed by modifying the data or by utilizing analytical models (Alexander 2023). A few common approaches involving data side strategies are discussed below.

Listwise Deletion

One straight-forward approach is to drop observations with missing data, also known as, listwise deletion (Alexander 2023). This approach involves removing any cases with missing values in any of the variables of interest from the analysis. This approach is relatively easy to implement, however, the drawbacks of this approach can include substantial reduction in the sample size if missing data are common. The reduced sample size results in a loss of statistical power and reduced efficiencies of data use. If the data are not MCAR or MNAR, listwise deletion can introduce much bias into the estimates (**CARMA?**).

Despite these limitations, listwise deletion can still be a valid choice when the proportion of missing data is small and can be assumed to be MCAR. Nonetheless since MCAR is rather rare, it is important to consider the nature of the missingness before implementing this approach.

Pairwise Deletion

A second strategy is pairwise deletion. The analysis is conducted using all available data pairs for each variable pairing. If some data points are missing for a specific variable, the missing cases are only omitted when calculating statistics that involve that variable (Newman 2014). For example, consider a survey that collects responses for social media usage, physical activity, and school performance. If a response is missing the social media usage data, it will still be included in the analyses involving physical activity and school performance but omitted from the analysis concerning social media usage.

Pairwise deletion enables the use of all available data, which can enhance the statistical power of an analysis. It is also considered more flexible when dealing with datasets that exhibit various patterns of missingness, facilitating analyses that might not be possible with listwise deletion. However, pairwise deletion can result in varying numbers of observations being used in different parts of the analysis. Additionally, similar to listwise deletion, pairwise deletion may introduce bias into the estimates if the data are not missing completely at random (MCAR) (Newman 2014).

Single Imputation

A third strategy is to impute the mean of observations without missing data. It refers to the method to calculate the mean value of the variable with missing data based on all the available, non-missing observations. This mean is subsequently used to fill in for the missing values in the original dataset (Zhang 2016).

This approach preserves the sample size compared to the listwise deletion method. However, it may reduce the variability in the data since the mean does not add new information to the dataset. If the data is not MCAR, mean imputation can also introduce bias. In addition, it can affect the correlations between variables, making them appear stronger or weaker than they are.

Multiple Imputation

Another data-based strategy is multiple imputation. In multiple imputation, a set of plausible values for missing data is generated, not just once but multiple times, creating several complete datasets (Little and Rubin 2019). These values are usually predicted based on the relationships observed in the rest of the data. Each of the completed datasets is then analyzed using statistical procedures, as if there were no missing data. The results from these multiple analyses are then combined to produce estimates and confidence intervals that reflect the missing data uncertainty.

Multiple imputation has several advantages over the other techniques. It provides a more accurate reflection of the uncertainty due to missing data than single imputation because it

takes into consideration the variability between the different complete datasets. As a result, it introduces less bias into the estimates when the missing data are MAR or MCAR. However, multiple imputation also requires strong statistical expertise to implement correctly. It also is more computationally intensive than single imputation methods (Newman 2014).

Conclusion

Missing data is a long-standing issue for researchers, and each study presents a unique challenge. Researchers should address the problem with careful thought (Alexander 2023), choosing their methods based on the amount, pattern, and the underlying mechanism of missing data. Newman suggested that practitioners in fields such as psychology and management particularly benefit from employing techniques like multiple imputation and other model-based strategies (Newman 2014). Mitigating the effects of missing data goes beyond mere statistical analysis. It is a crucial part of upholding the integrity of scientific research.

References

Alexander, Rohan. 2023. "Telling Stories with Data with Applications in r and Python." https://tellingstorieswithdata.com/03-workflow.html.

Little, Roderick J. A., and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data.* 3rd ed. John Wiley & Sons.

Newman, Daniel A. 2014. "Missing Data: Five Practical Guidelines." *Organizational Research Methods* 17 (4): 372–411.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–92. https://doi.org/10.1093/biomet/63.3.581.

Zhang, Z. 2016. "Missing Data Imputation: Focusing on Single Imputation." *Annals of Translational Medicine* 4 (1): 9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38.