Discuss Missing Data*

Heyucheng Zhang

24 February 2024

Table of contents

1 Introduction

Data that is intended to be in a dataset but isn't is known as missing data. The absence of details can take many different forms, ranging from a single missing value in a survey to whole population data sets missing. The integrity of statistical studies has been harmed, which may result in conclusions that are not truly representative of the population of interest, results that are biased, and decreased statistical power. Research and analysis in a number of fields, including economics and data sciences, suffer greatly from the existence of missing data. Following a reading of several chapters from "Telling Stories with Data"(Alexander 2023) about missing data, the study focuses on what defines missing data and what can be done about it. For researchers hoping to obtain reliable and precise conclusions from their data, it is essential to fully understand the nature of missing data, its effects, and solutions. The study aims to provide researchers with the knowledge and resources they require to reduce the effects of missing data, thereby improving the reliability and accuracy of their findings. This is achieved by identifying three types of missing data and exploring various solutions for handling missing data, from simple deletion methods to advanced imputation techniques.

[*] Code and data are available at: https://github.com/heyuchengzhang/Discuss-missing-data. Reviewed by Jiwon Choi, and Updated the paper based on the feedback.

2 Main body

Based on Gelman, Hill, and Vehtari (2020), missing data in statistical analysis is divided into three different categories. They are Missing Completely At Random (MCAR), Missing at Random (MAR) and Missing Not At Random (MNAR). Each has unique difficulties that call for particular strategies for efficient management. Missing Completely at Random is the least problematic scenario because it doesn't bias study results, but it is relatively rare in real-world data. MCAR occurs when the probability of data being missing is equal across all observations, making it unrelated to both observed and unobserved data. MAR describes scenarios in which the absence of data can be made up for using the information in the dataset, however requiring additional statistical techniques to produce unbiased results. Missingness is associated with observed data but not the missing data itself. The most challenging scenario is known as MNAR, which occurs when the missingness is caused by factors related to the unobserved or missing data itself. If particular methods are not used to address this systematic pattern in the missing data, it may result in significant analysis bias. There are various sources of missing data, and each one adds to dataset gaps that make analysis difficult. One frequent reason is non-response, where participants choose not to respond to some survey questions or leave research early, leaving behind partial responses(Newman 2014). In continuous investigations, attrition causes this problem by gradually producing increasingly incomplete datasets due to participant dropout. Inaccuracies throughout the data gathering or recording process result in missing or erroneous data points, adding another level of complexity to data entry errors. In addition, instrument failure may happen, in this case, the instruments or apparatus applied to collect data fail, interfering with the procedure and producing insufficient data.

To minimize the effect on the results of studies, each of the above sources needs to be handled properly and with a lot of thought. Dropping observations with missing data or complete case analysis, which removes all records with missing values, is a simple method that can greatly reduce the size of the dataset and possibly introduce bias, especially if the data isn't missing completely at random (MCAR). To drop the observations with missing data, we can use mean()(Alexander 2023). Imputation is an additional technique that seeks to fill in the gaps with estimated values based on the information already present in the dataset. Simpler methods like mean, median, or mode can be applied, as an advanced ones like multiple imputation, which generates, analyzes, and then aggregates multiple filled-in versions of the dataset to better account for the uncertainty of missing data(Gelman and Hill 2007). We can implement multiple imputation with mice() from mice(Alexander 2023). The method chosen to deal with missing data mostly depends on the type of missingness and the particular study environment.

3 Conclusion

Missing data is an unavoidable research issue that needs to be carefully considered and handled properly to ensure the validity and reliability of study findings. The negative impact of missing data on research can be reduced by researchers by identifying the different kinds of missing data and using suitable techniques to resolve them. As statistical techniques advance, so do the strategies for handling missing data, providing more solutions to this common problem.

Reference

Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in r*. Chapman; Hall/CRC. https://doi.org/10.1201/9781003229407.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. 1st ed. Cambridge University Press.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge University Press. https://avehtari.github.io/ROS-Examples/.

Newman, Daniel. 2014. “Missing Data: Five Practical Guidelines.” *Organizational Research Methods* 17 (4): 372–411. https://doi.org/10.1177/1094428114548590.