What is missing data and what should we do about it?*

Yaning Jin

March 4, 2024

The phenomenon of missing data in research encompasses the absence of data points that were intended to be collected but were not. This issue is pivotal because it can significantly impact the validity and reliability of research findings. Identifying the patterns and mechanisms of missing data—MCAR, MAR, and MNAR—is crucial for selecting the appropriate analytical approach to mitigate its effects. This paper explores the strategies for handling missing data, including deletion methods, imputation techniques, and likelihood-based approaches, emphasizing the importance of choosing a method that aligns with the data's missingness mechanism to ensure robust research outcomes.

Table of contents

*Code and data are available at:https://github.com/Yuki010305/Miniessay-8.git

1 Introduction

In research, missing data presents a significant challenge, potentially skewing results and leading to incorrect conclusions. It occurs for various reasons—ranging from non-response in surveys to errors in data collection—and its nature significantly influences analytical outcomes. Understanding its implications is crucial for researchers aiming to maintain the integrity of their findings. This introduction emphasizes the importance of recognizing the different mechanisms through which data can be missing and the subsequent selection of appropriate handling techniques, setting a foundation for a comprehensive discussion on strategies to mitigate the impact of missing data on research.

2 Analysis

The first step in addressing missing data involves recognizing its potential sources and mechanisms. MCAR occurs when the probability of missing data on a variable is independent of any observed or unobserved data.(Baraldi and Enders 2010) MAR happens when the missingness is related to the observed data but not the unobserved missing data. MNAR is when the missingness depends on the unobserved data. Understanding these mechanisms is essential because they guide the selection of appropriate handling techniques. Deletion methods, such as listwise or pairwise deletion, are simple but can lead to biased results if the data are not MCAR. Imputation methods, including mean substitution, regression imputation, and multiple imputation, attempt to fill in missing values based on the observed data.(Baraldi and Enders 2010) Multiple imputation and maximum likelihood estimation are advanced techniques that provide more reliable results under MAR and MNAR conditions by incorporating uncertainty about the missing data.

3 Discussion

Choosing the correct method for handling missing data is critical to maintain the integrity of research findings. Deletion methods can significantly reduce the sample size, leading to a loss of statistical power. Single imputation methods introduce bias and underestimate variability since they treat imputed values as true values. In contrast, multiple imputation accounts for the uncertainty around missing data by creating several imputed datasets, leading to more accurate standard errors and confidence intervals. Maximum likelihood estimation, particularly with the expectation-maximization algorithm, offers a flexible approach to estimate model parameters directly without imputing missing values. (Baraldi and Enders 2010) However, these advanced techniques require assumptions about the data and missingness mechanism, underscoring the importance of rigorous preliminary analysis.

4 Conclusion

Missing data poses a significant challenge in research, potentially undermining the validity and reliability of findings. Identifying the mechanism of missingness is fundamental to choosing an appropriate handling technique. While traditional methods like deletion and single imputation are straightforward, they often introduce bias and reduce statistical power. Advanced methods such as multiple imputation and maximum likelihood estimation offer robust alternatives that better account for the uncertainty of missing data, provided the assumptions about the missingness mechanism are met. Ultimately, carefully addressing missing data through appropriate techniques is essential for achieving accurate and reliable research outcomes.

Baraldi, Amanda N, and Craig K Enders. 2010. "An Introduction to Modern Missing Data Analyses." *Journal of School Psychology* 48 (1): 5–37.