Investigation of Missing Data: Strategies and Implications*

Dingning Li

2024-03-05

Introduction

In research, missing data is a common and prevalent issue that can affect statistical analysis and lead to misleading conclusions. It occurs when information is not observed or recorded, influencing the dataset's integrity. There are three types of missing data situation: Missing Completely at Random(MCAR), where missing data is unrelated to any variables; Missing at Random(MAR), where missing data is related to other observed variables but not the missing data itself; and Missing Not at Random(MNAR), where the missing data is related to the value of the missing data(Baraldi and Enders 2010). Understanding complex mechanisms is important for applying appropriate method to solve the issues of missing data on research findings.

Types of Missing Data

Missing Completely Random Data(MCAR)

MCAR occurs when all observations are missing with the same probability. it implies that there is no systematic difference between missing and observed data, making it the least problematic type because it does not introduce the bias associated with observed or missing data. MCAR is a rare case of missing data.

Missing at Random Data(MAR)

MAR occurs when the probability of missing data is related to the observed data and not to the missing data itself. These data are not completely random missing, the mechanism behind their loss can be explained in terms of other observed variables.

Missing Not at Random Data(MAR)

MNAR refers to data missing due to reasons related to the missing data itself or unobserved variables. This type of data missing can lead significant bias, as the absence of data which is related to its unseen values.

Strategies for Handling Missing Data

In research, handling missing data efficiently is essential for the integrity of statistical analyses and reliability of the results. Various strategies have been introduced to address this challenge, including simple methods that exclude missing data to more complex and sophisticated mechanisms. Each method has its strengths and limitations, and it is crucial to make the choice of method based on the nature of missing data and the research objectives. In the following discussion, I will introduce these approaches and determine the optimal approach for addressing missing data.

*Code and output of pdf are available at:https://github.com/iamldn2002/Quiz-or-Tutorial/tree/main/missing_data

Deletion Method:

Deletion strategy is the most basic and traditional method to address missing data issue. This approach refers to removing data points or record that contain missing values, providing a complete data set finally. But it can also lead to significant data loss and cause potential bias if the missing data is not completely random.

Single Imputation

A more prevalent strategy is the process of replacing missing data with substituted values to allow for complete data analysis. Single imputation involves filling in each missing value with a single, specific value, such as mean, median, or mode of the available data. They are simple to implement but can underestimate variability and introduce bias.

Multiple Imputation

A more sophisticated technique used to handle missing data fills in each missing values several times, creating multiple complete datasets. These datasets are then analyzed separately, and the results are combined to produce estimates that reflect the uncertainty due to the missing data. The technique is suitable for various data types and missing data patterns, but need for appropriate imputation models.

Case Studies and Applications

Real World Example: Investigate public satisfaction with public facilities. I create a case study simulating public satisfaction with public facilities, where participants rated their satisfaction on a scale from 1 to 10 and provided their gender and age. This simulation included missing data to mimic real-world issues researchers might face. In the case study, I will use (R Core Team 2020) and (van Buuren and Groothuis-Oudshoorn 2011) packages to simulate data and impute data into the dataset.

```
##           Method     Mean Median
## 1       Original 5.771591      6
## 2 After Deletion 5.771591      6
```

```
##   Statistic Original After_Imputation
## 1      Mean 5.771591         5.771591
## 2    Median 6.000000         5.771591
```

```
## Warning: Number of logged events: 1
```

```
##                Method     Mean Median
## 1            Original 5.771591      6
## 2 Multiple Imputation 5.818000      6
```

Upon comparing the three methods of handling missing data, the deletion method did not change the mean or median, suggesting that the missing data were likely MCAR, not biasing the sample's central tendency. The single imputation method, which filled missing values with the overall mean, maintained the original mean but slightly decreased the median, indicating that mean imputation could slightly distort the distribution of the data. Multiple imputation, a more sophisticated approach, provided a slightly higher mean than the original, which might be due to the imputation model capturing underlying patterns in the data to estimate the missing values.

Each method has implications for data analysis: deletion can reduce sample size, single imputation may underestimate variability, and multiple imputation generally provides a more robust estimate accounting for the uncertainty of missing data. The choice among them should be guided by the missing data mechanism and the research context.

Conclusion

In cobclusion, the exploration of missing data reveals its pervasive impact on research integrity and the need for careful treatment. Key strategies—deletion, single imputation, and multiple imputation—each play a distinctive role, tailored to the nature of the missingness. As we advance, the pursuit in missing data research is directed towards refining imputation techniques, leveraging machine learning algorithms, and enhancing our understanding of the biases introduced by data missing. These developments are expected to enhance the robustness of statistical analysis in the face of incomplete data.

Acknowledgment

I am grateful to my peer, Yunshu Zhang, for her insightful suggestions that have significantly enhanced the quality of this paper. Her thoughtful feedback have been helpful in strengthening the analysis presented.

Reference

Baraldi, Amanda N., and Craig K. Enders. 2010. "An Introduction to Modern Missing Data Analyses." *Journal of School Psychology* 48 (1): 5–37. https://doi.org/10.1016/j.jsp.2009.10.001.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.