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

Yetao Guo

2024-03-04

*Code and data are available at: https://github.com/yetaoguo/miniessay8.git

Missing data frequently poses a severe threat to the validity and reliability of the statistical analysis used to produce the study's conclusions. Researchers develop different methods and frameworks to handle the missing data effectively, therefore ensuring data analysis is done with intercity, thus drawing insights from the recent studies, which include Carpenter and Smuk (2021), Leyrat et al. (2020), Enders (2022) and lastly Little and Rubin (2019), we explore the types and strategies to handle missing data for its management.

The three primary different missing data are MCAR, MAR, and MNAR.First of all, Missing Completely at Random(MCAR) refers to the missingness of data points. Both observed and unobserved data are unrelated in this scenario. Thus, it suggests that the probability is the same for the missingness of all observations. For example, if the responses of a survey are lost due to technical issues during data entry, this missingness would be concluded as MCAR. Secondly, Missing at Random (MAR) indicates that the seen data determines the likelihood of a missing event at random, not the unobserved data per se. Observed variables, once accounted for the missingness, are arbitrary. For instance, individuals who have higher income in a survey are less likely to disclose how much they are paid, but it can be explained in other patterns explained by other variables.An example is age or the level of education which is seen. Thirdly, Missing Not at Random (MNAR) shows that When the missingness to the unobserved data itself occurs, then MNAR happens even after the observed variables are accounted for. Bias in the analysis can be introduced by this type of missingnmess if not appropriately handled. For instance, people who have a higher level of depression during a survey are likely not to report the status of their mental health. At the same time, even after considering the variables observed, the pattern persists, like gender and age.

Various approaches have been suggested for dealing with missing data, according to Leyrat et al. (2020). Thus, to address the lost data, careful consideration is required to ensure the analysis is done with integrity. The following strategies for handling missing data can be employed: Firstly, the method of Complete Case Analysis (CCA) is known as Listwise deletion. This method excludes any observation with missing values from the analysis. Since CCA is more accessible to implement, it may lead to results that are biased if the missing data is not entirely random. Therefore, statistical power may also be reduced by discarding valuable information. Next, Imputation Techniques are also applied to deal with missing data. Imputation alternates calculated values for missing values based on observable data. Mean, Mode, Regression, and Multiple imputation to provide data analysis are the most expected forms. Thirdly, Model-Based Methods will be used to estimate missing values in data analysis of the findings in a survey; these model-based methods are used to provide an observed data representative. The most ordinary algorithms used for the missing data in this method include expectation maximization (EM).Besides, the Weighting Schemes method shows that the existing bias is addressed, particularly in the performed research survey, by assigning weights to the observation of the missing data based on the likelihood of missingness; thus, it can enhance the sample representation of the findings and improve it overall. Last but not least is the method of Sensitivity Analysis. It evaluates the viability of missing data of different hypotheses about the mechanisms underlying is crucial. Sensitivity analysis, which

includes systematically varied assumptions, involves evaluating how correspondingly findings alter.

Researchers such as Carpenter and Smuk (2021), Leyrat et al. (2020), Enders (2022), and Little and Rubin (2019) assist in integrating the findings to give a more thorough toolset for handling missing data. Using suitable handling techniques and a lost data design, researchers can lower preferences and guarantee the validity of their discoveries. Consequently, sensitivity analysis allows researchers to examine the robustness of their results, expanding the study's findings' conviction in their reliability.

In conclusion, better data is vital for statistical analysis, which creates barriers, but researchers maintain helpful management techniques. Understandings such as proper handling techniques, imputation model-based methodologies, sensitivity analysis, and weighting analysis are removed from existing studies; as a result, reliable research findings may be developed, and researchers can handle them safely when guiding missing data. Because data collection is changing, researchers must be on the lookout to ensure authenticity and integrity. Concentrated study outputs necessitate detailed application tactics and understanding the means underlying missing data when dealing with difficult-to-find missing data.

References

Carpenter, J. R., & Smuk, M. (2021). Missing data: A statistical framework for practice. Biometrical Journal, 63(5), 915-947.https://doi.org/10.1002/bimj.202000196

Clémence Leyrat, Carpenter, J. R., Bailly, S., & Williamson, E. J. (2020). Common Methods for Handling Missing Data in Marginal Structural Models: What Works and Why. American Journal of Epidemiology, 190(4), 663–672. https://doi.org/10.1093/aje/kwaa225

Enders, C. K. (2022). Applied missing data analysis. Guilford Publications. https://books.google.co.ke/books?h

Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley&Sons. https://books.google.co.ke/books?hl=en&lr=&id=BemMDwAAQBAJ&oi=fnd&pg=PR11&dq=mis
bzujI4Gqipg9rFM&redir_esc=y#v=onepage&q=missing%20data&f=false

peer review for YANING JIN