What to do about missing values*

Ahnaf Alam

February 21, 2024

Table of contents

1 General idea **1**

2 Missing completely at random **2**

3 Missing at random **2**

4 Missing not at random **2**

5 What to do about missing data **3**
5.1 Listwise deletion 3
5.2 Pairwise deletion 3
5.3 Single imputation 3
5.4 Maximum likelihood 3
5.5 Multiple imputation 4

6 Conclusion **4**

Citations **5**

1 General idea

Missing data points are a regular feature in almost all datasets. Researchers' defining task is figuring out the source of the missing data, as these can lead to bias and error (Newman 2014). Missing data can arise for many reasons; however, we can categorize missing values into three distinct categories:

*For further information, please visit https://github.com/AhnafAlam1/missing-values

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

The subsequent paragraph will discuss these categories of missing values, their features, reasons, and what researchers can do to mitigate the issue of missing data.

2 Missing completely at random

This situation arises when data points are missing and independent of other variables (Alexander 2023). In layman's terms, this is when data points are missing randomly and unrelated to different factors or variables in the data set. This coincidental situation implies the data set in hand still represents the broader population. Amongst the categories of missing data, this form of missing data is the only one that is entirely random. However, this form of missing data is also scarce as this type of data requires researchers to make "untenable assumptions" (Graham 2009).

3 Missing at random

MAR differentiates from MCAR in that missing values are related to one of the other variables in the data set. Take, for example, a situation where a researcher is interested in the differing Indian caste system's opinion on the country's current federal government. In this situation, we will have MCAR if we have missing values for the opinion variable for lower castes compared to higher castes. MAR is a systematic missing data mechanism (Newman 2014), unlike MCAR.

4 Missing not at random

MCAR occurs when missing data is related to an unobserved or latent variable or when missing data is associated with the missing variable itself. Suppose you are collecting data on a patient's weight in a medical clinic. You ask patients to self-report their weight to the researcher every two weeks. Now, imagine individuals with higher weight levels may feel insecure about sharing their weight details and may need to report their weight more regularly. In this situation, higher weights are related to missing data for weight. It is often impossible to determine if data is MNAR, as this requires the researcher to compare observed values to missing values, and the researcher needs access to missing values (Newman 2014).

5 What to do about missing data

There are five common ways to deal with missing values in data sets (Newman 2014). They are:

1. Listwise deletion
2. Pairwise deletion
3. Single imputation
4. Maximum likelihood
5. Multiple imputation

5.1 Listwise deletion

The basic idea behind listwise deletion is to delete all cases where data is missing and proceed with the analysis. The apparent issue with this type of mechanism is we risk reducing the sample size significantly, especially if there are a lot of missing data points. This is bad as lowering the count of observations means the power for statistical significance decreases, and this could lead to biased estimates, particularly for MAR and MNAR.

5.2 Pairwise deletion

Pairwise deletion omits data based on variables used in the analysis. In other words, the analysis may be completed based on a subset of data depending on where the missing value is (Moran 2024). For example, suppose person X has missing data for income but has complete data for race, height and weight. In that case, X will be included in all analyses that involve race, height, and weight but excluded from analyses that include income. This, again, could lead to biased estimates, especially for MAR and MNAR.

5.3 Single imputation

Single imputation involves filling in the missing value using some form of estimate, usually being mean, median or mode. However, doing this leads to underestimation of variance and this methods ignores relationship with other variables (Zhang 2016).

5.4 Maximum likelihood

This method involves directly estimating parameters of interest from incomplete data and proceeding with analysis based on this (Newman 2014). This form of estimation relies on probability when estimating missing data. It considers the likelihood of scenarios and estimates

based on what is more probable given the existing data. One of the positives of this mechanism is that it is unbiased under MCAR and MAR (Newman 2014).

5.5 Multiple imputation

This mechanism draws "repeated simulated datasets from a posterior distribution defined by the missing data conditional observed data" (Pepinsky 2018). This method generates several potential solutions for missing data, conditional on what you already know from the data. This is a convenient model as it improves as you add more variables to the model, and further, it is unbiased under MCAR and MAR (Newman 2014).

6 Conclusion

Missing data can arise for various, and there are multiple methods to deal with. Researchers prefer some as analysis leads to unbiased estimates. However, in most cases, the missing value type dictates what mechanism is needed to deal with this issue. Table 1 provides a summary of all the different mechanisms used to deal with missing data, along with their benefits and drawbacks, based on Newman (2014).

Note: The table is created using R (R Core Team 2023), a statistical software, along with libraries tidyverse (Wickham et al. 2019), kableExtra (Zhu 2021) and knitr (Xie 2023).

Table 1: Table display major benefits and solutions for each of the missing data mechanism. It is based on Newman (2014).

Potential solutions	Benefits and drawbacks
Likewise Deletion	Biased under MAR and MNAR
Pairwise Deletion	Biased under MAR and MNAR
Single Imputation	Biased under MCAR
Maximum Likelihood	Unbiased under MAR and MCAR. The model improves as you add more variable
Multiple Imputation	Unbiased under MCAR and MAR. Also improves as you add more variables to the model. However, the model gives slightly different estimate each time

Citations

Alexander, Rohan. 2023. "Exploratory Data Analysis." *Telling Stories with Data - 11 Exploratory Data Analysis*. https://tellingstorieswithdata.com/11-eda.html#missing-data.

Graham, J. W. 2009. "Missing Data Analysis: Making It Work in the Real World." *Annual Review of Psychology* 60: 549–76.

Moran, Melissa. 2024. "Handling Missing Data: Listwise Versus Pairwise Deletion." *Statistics Solutions*. https://www.statisticssolutions.com/handling-missing-data-listwise-versus-pairwise-deletion/#:~:text=Pairwise%20deletion%20omits%20cases%20based,on%20where%20values%20are%20missing.

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

Pepinsky, Thomas B. 2018. "A Note on Listwise Deletion Versus Multiple Imputation." *Political Analysis* 26 (4): 480–88.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

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.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.