Missing data's meaning and how we should deal with it.*

Yunzhao Li

March 5, 2024

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

What is missing data?

Missing data occurs frequently in research, and how well it is handled will affect the we conclusions drawn. Methodologist William Shadish, stated in (Azar 2002) has identified missing data as "one of the most important statistical and design problems in research" due to its widespread nature. (Baraldi and Enders 2010)

A definition given by (Little and Rubin 2020) is " Missing data are unobserved values that would be meaning - ful for analysis if observed ; in other words , a missing value hides a meaningful value ."

There are three main categories stated in (Alexander 2024):

"1.Missing Completely At Random

2.Missing at Random

3.Missing Not At Random."

And we often need to deal with them using different approaches.

*Code and data are available at: https://github.com/yunzhaol/missing_data.git

What should we do about it?

According to (Alexander 2024), we mainly have these 3 methods to handle them:

"1.Drop observations with missing data.

2.Impute the mean of observations without missing data.

3.Use multiple imputation."

We can use R's (R Core Team 2023) mean() function to remove the observations with missing data. By default, observations with missing values are not included in the calculations. To estimate the mean, we create an alternative dataset that omits any observations that have missing data. Next, we calculate the mean for the related column in this new dataset. This mean is then used to fill in the missing values in the initial dataset. Multiple imputation involves generating several plausible datasets, performing analysis on each, and then combining the results, often by taking the average. (Baraldi and Enders 2010)

Running simulations to eliminate existing observations and then applying different alternatives can enhance our comprehension of the compromises we encounter. Regardless of the decision taken, since seldom is there a straightforward solution, we have to make an effort to record and share the actions taken, and investigate how varying decisions impact later predictions.(Alexander 2024)

Sometimes, a variable with specific values is used to encode missing data. For example, though R offers the option "NA," numerical data is occasionally input as "-99" or, in the event that it is missing, as a very big number like "99999999". Here introduced =three types of known missing data:

""888": "Asked in this wave, but not asked of this respondent"

"999": "Not sure, don't know"

".": Respondent skipped"

Graphs and tables are often very helpful for the whole procedure. (Alexander 2024)

Furthermore, there are several traditional and modern techniques stated in (Baraldi and Enders 2010). "The most common of traditional techniques include deletion and single imputation approaches (Peugh & Enders, 2004)."

"Maximum likelihood estimation and multiple imputation are considered "state of the art" missing data techniques (Schafer & Graham, 2002) and are widely recommended in the methodological literature (Schafer and Olsen, 1998, Allison, 2002, Enders, 2006)"

Also, according to (Little and Rubin 2020),we have another approach named finding the Missing Values through Iteration: It was suggested by Hartley (1956) that a general noniterative method for predicting one missing value be applied iteratively for several ones. Three distinct trial values are used in place of the missing value in the technique for one missing value, and

the residual sum of squares is computed for each trial value. The minimizing value of the one missing value can then be determined because the residual sum of squares in that one missing value is quadratic.

Credit

This work is updated based on Wentao Sun's feedback.

Reference

Alexander, Rohan. 2024. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com.

Azar, B. 2002. "Finding a Solution for Missing Data." *Monitor on Psychology*.

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.

Little, Roderick J. A., and Donald B. Rubin. 2020. *Statistical Analysis with Missing Data*.

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