The Mystery Of Missing Data*

Xincheng Zhang

March 4, 2024

Table of contents

Introduction **1**

What Is Missing Data? **2**
Three Types of Missing Data 2

What Should You Do About It? **2**
Deletion 3
Single Inputation 3
Reweighting & Averaging The Available Items 3

Conclusion **4**

References **4**

Introduction

Data plays a dominant role in the field of statistics because data can give statisticians many kinds of information. Different data types represent different functions. For example, numeric type data can give the user some information related to numbers, and text type can explain some different text associations. However missing data is a special type of data, which means that some variables and observations in the data set are missing or incomplete. Missing data is a very common phenomenon and creates problems to varying degrees in various teaching and cultural fields, such as psychology.(Baraldi and Enders (2010))

*Files are Availiable at

What Is Missing Data?

Specifically, in modern mainstream concepts, types of missing data can be roughly divided into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). These three different categories are explained according to probability in statistics, and the content explained is the three classifications of the main reasons for missing data. Next, I will introduce these three different types of missing data.(Baraldi and Enders (2010))

Three Types of Missing Data

For MCAR, it represents the situation when the probability of missing data and the required data are not related. To be more specific, when the missing data is not related to the study variable and does not have any impact, it is classified as MCAR. For example, when a statistician calculates the annual salary of graduates, all relevant columns are "age", "major", "school", and "region". During the process of collecting data, some missing variables in the column of "gender" were found, which will be called MCAR. Because "gender" is not in the required column, it will not have any impact on the results in these statistics.

In terms of missing data in the MAR classification, when this data type is missing, it will have an indirect relationship with the variables under study. In other words, missing data related to other required and collected variables in the analysis model are classified as MAR. For example, when statisticians analyze the psychological burden of the experimenters, the economic status (generally divided into low/medium/high) will indirectly affect the test results of each person's psychological burden. Because a large number of people still have psychological pressure on the financial burden.

Finally, there is MNAR, which represents the direct relationship between the missing data and the object under study. Missing data is classified as MNAR when it affects the statistician's analysis of the overall data model. When analyzing students' learning pressure, if there is missing data in the column number of courses per semester, it will have an impact on the entire survey because the number of courses is very decisive and relevant to learning pressure.

The above is the explanation and classification of missing data in statistics.

What Should You Do About It?

Based on the extensive use of statistics and attempts to deal with missing data, researchers and statisticians have summarized the two main techniques in modern times, namely deletion and single imputation approaches. Joseph L. Schafer additionally mentioned the method of reweighting and averaging the available items (Schafer and Graham (2002)). These methods

can make the final processed database more reasonable and enhance credibility by targeting the phenomenon of missing data that affects statistical results.

Deletion

The deletion method is the most common way to deal with missing data. The specific method is just like its name, it deletes the row with missing data. This simple method has many advantages. First of all, it is simple to operate. Users only need to set code instructions to select and discard all data with missing values. Secondly, this method can generate a complete table at the end and contain the corresponding type of data in each entry, thus ensuring the integrity of the data set. On the contrary, the shortcomings of the deletion method will be exposed when there are too many missing data because when most rows have missing values, this method will lead to discarding most of the data we collected, thus affecting the sample size and greatly reducing the data reliability. This method is best used in the missing type of MCAR because deletion does not affect the overall statistical results.

Single Inputation

Another method frequently used by researchers to deal with missing data is called Single Inputation. Its main step is to replace missing data with other appropriate values. The two different replacement data are the mean values, and values that are predicted by the regression equation. The mean is to average the existing data except for missing data and replace each missing value with this average. The linear regression equation can also be used as one of the methods to estimate missing data. A regression equation is established through the existing data and the complete variables are used as predictor variables to estimate the missing data. This approach can ensure that the gap with the actual results is narrowed while ensuring the sample size, so it is widely used.

Reweighting & Averaging The Available Items

The reweighting method refers to adjusting the weight assigned to the observation value. For example, it can make the existing data have a greater decisiveness and proportion in the final result to reduce the impact of missing data on the result. However, this method can only be used for data measurement errors or inaccurate data measurement due to technology and cannot be used when the data is completely missing. This is because when data are missing, the weight cannot be reduced as this would ignore the importance of the single data.

Averaging The Available Items is relatively not widely used because this method lacks certain recognition and is extremely limited. It can only be used to reintegrate some highly correlated variables in the data to make the variables in this new large set replaceable to replace the missing data with other data. This method should not be used in MCAR because it will cause

bias. Specifically, when we replace some missing data (MCNR) that are highly correlated with the results with some relatively uncorrelated data, it will cause bias. Undesirable error.

Conclusion

The above introduces the definitions and three main classifications of missing data and explains the advantages and disadvantages of modern methods for dealing with missing data. The four seemingly useful and reliable methods mentioned above were actually denied by Judi Scheffer through observations and experiments. The method of substituting average values into missing data actually destroys the variance structure of the original data and changes it into an unnatural data set. For missing data other than MCAR categories, the deletion method is also considered wrong because it will seriously reduce the sample size. Thus, for statisticians and researchers, it is most reasonable to avoid data missing as much as possible and to choose the least destructive method according to the type of data missing.(Scheffer (2002))

References

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

Schafer, Joseph L, and John W Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2): 147.

Scheffer, Judi. 2002. "Dealing with Missing Data."