Missing Data

Advanced Data Analysis: Independent Study Activity 1

Orville D. Hombrebueno
odhombrebueno@nvsu.edu.ph
PhD in Science Education: Mathematics Student
Saint Mary's University
Bayombong, Nueva Vizcaya

February 12, 2020

Instruction: Perform the following tasks relative to the reality or inevitability of "missing data" or "missing values".

1. Explain the nature of missing data or missing values and the importance of coding and mechanically treating missing values. Elaborate on the serious implications of not coding missing values to the integrity of a research.

   The importance of dealing with missing was discussed by Ho (2014) in the following:

   > Encountering missing data in research is part and parcel of data analysis. The problem arises for a number of reasons; examples include subjects dropping out of an experimental condition because of boredom/fatigue; refusal to answer a particular question on a survey questionnaire; death; and refusal to participate in the post-test of a longitudinal study. Whatever the reasons, analysis of a "missing data set" can be problematic if the pattern of missing data is non-random. Any statistical results based on such data would be biased to the extent that the variables included in the analysis are influenced by the pattern of non-randomness of the missing data. According to Tabachnick & Fidell (2001), if only a few data points ($\leq 5\%$) are missing in a random pattern from a large data set, the problems are less serious and almost any procedure for handling missing values yields similar results. Thus, when screening data, it is important to identify the pattern of missing data. (p. 23-24)

   There are different user-defined coding schemes that can remind us why or how the data is missing. Missing data when properly coded can be easily summarized. This summary will contribute to understanding the data gathering procedure and the respondents. One can see if the data gathering procedure is effective. Also, one will understand how respondents answer given questions. These insights can be used to back-up findings in the study. Further more, one can easily understand the limitation of the study when coding and treatment of missing data is properly done and discussed in the study. Otherwise, one might

be claiming a finding that has really no basis since the coding and handling of missing data is not properly done (e.g. more than 50% of the data are missing). The result of analysis of data with missing data is different from data without missing data; the one without missing data is more reliable.

2. Explain the following labels for missing data/value (e.g. MANN = Missing and not needed, when the characteristic of interest or condition is absent and not necessary).

   2.1. NA

   Not answered. This is a label for an item that has no entry. The respondent did not answer the item for some reason. One might don't know the answer but we don't know this (*Chapter Four: Univariate Statistics*, n.d.).

   2.2. NAP

   Not appropriate. This label is for items that have entries but are not appropriate for the given questions. Example: the question is asking for monthly income but the respondent entered age. The answer is different from what is expected. The researcher might not figure out that it is age (*Chapter Four: Univariate Statistics*, n.d.).

   2.3. NAV

   I have not yet encountered this user defined code. I will use it as a code for "not a value". For instance, I am asking for height in millimeters but I got height in inches. I can convert it if I want but, I can also use the code as a category label to consider compared to the rest.

   2.4. DK

   Don't know. The respondent placed "I don't know." as answer to an item (*Chapter Four: Univariate Statistics*, n.d.).

   2.5. DC

   I can use this to label "don't care". The respondent simply wrote "I don't care." as answer. This is if I care about those who don't care to answer the question.

2.6. Refused

The respondent refused to answer the question since the information is confidential. It might be for other reasons. Refusing to answer is implied by the respondent (*Chapter Four: Univariate Statistics*, n.d.).

2.7. NO

No observation. I will use this for time series data where I find no observations available for certain data points in time.

2.8. NOP

Not an option. The answer is not in the list of options.

2.9. DR

This can mean "did not respond". This can be a label for those who do not have a post test scores in a longitudinal study.

2.10. NS

This can mean "not sure". It can be considered as a category.

2.11. Illigible

This are answers that are writen in a way that one cannot understand. The researcher cannot understand the answer because the portion in the paper where the answer was writen got wet for some reason.

To be honest I am not aware of these labels: NAP, NAV, DK, DC, Refused, NO, NOP, DR, NS, Illegible. I don't even know how to use them since I have not yet done research wherein I gather data through questionnaires. I usually use secondary data for research and usually use NA for missing data. NA simply means it is missing.

3. Describe specific instances in your own particular discipline or area of research interest to illustrate (5) different types of the above missing values of data that may be gathered (e.g. MANN: In surveys about facilities for photography, darkroom is no longer necessary as photos can be easily developed using digital printing).

   NA: In any field when using questionnaires to gather data, some respondent simply don't enter a value for the item for some unknown reason.

   NAV: I have already discussed how I will use this label above.

DK: I can always use DK for "I don't know". It would be interesting to know in any study why some respondents answered "I don't know".

DC: "I don't know" is different from "I dont care". It would also be interesting to know in some studies why a respondent don't care in answering a certain question. This can occur in questions asking for opinions.

Illegible: I can always use this for answers that I cannont comprehend.

4. Propose three distinct ways of controlling or preventing the occurrence and undesirable sources of "missing" data/values.

   Proper administering of questionnaires can prevent missing data. Proctors should be very good in explaining each question. The proctor should also be checking missing values while administering the questionnaires by roaming around and looking at the answers of the respondents.

   There are different installed devices that gather data: devices that gather meteorological data, CCTV cameras, sensors, satellites, etc. Some rely on solar power. But power can be a problem. Not to mention wearing out of parts. Maintaining these devices properly can prevent missing data.

   As much as possible, make everything reproducible (Gandrud, 2017). You can gather data using online questionnaires. Data transfer from this interface to your laptop will result to 0 missing data. Unlike if someone will encode the answers; it is prone to error which can result to missing data.

5. Differentiate among the following in mechanical and statistical procedures dealing with cases with missing data:

   5.1. Listwise Deletion

   Listwise deletion is discussed by Ho (2014) below:

   > Cases with missing scores on any variable are excluded from all analysis. The effective sample size with listwise deletion includes only cases with complete records. An advantage of this method is that all analyses are conducted on the same

> number of cases. However, if missing observations are scattered across many cases, then deletion of cases can mean substantial loss of subjects. (p. 29)

5.2. Pairwise Deletion

Pairwise deletion is also discussed by Ho (2014) below:

> Cases are excluded only if they have missing data on variables involved in a particular computation. This means that the effective sample size can vary from analysis to analysis. This feature of pairwise deletion presents a potential problem for the statistical procedure of structural equation modeling as it can lead to the problem of a nonpositive definite matrix or singularity. (p. 29)

5.3. Mean Replacement

According to Ho (2014), mean replacement "involves replacing a missing score with the overall sample average. This method is simple, but it tends to distort the underlying distribution of the data, reducing variability and making the distributions more peaked at the mean".

6. In what special instances of missing data can sampled units be subjected for further relevant case analysis instead of listwise or pairwise elimination, or value imputation?

   Here is something interesting to share. In the book of Ho (2014), patterns of missing data was discussed.

   > Suppose two variables (X and Y) are collected. X has no missing data, but Y does have some missing data. *Missing at random* (MAR)–Missing data are termed MAR if the missing values of Y depend on X, but not on Y. Another way of stating this is that the missing values for Y can be explained by X in the data set. However, after accounting for X, the missing values of Y are random. An example can illustrate this. Assume that we know the marital status of respondents (the X variable), and we ask them about their income (the Y variable). Income is MAR if the probability of missing data

on income depends on marital status, but within each category of marital status (single, married, divorced) the probability of missing values for income is unrelated to the value of income. *Missing completely at random* (MCAR)–This is a higher level of randomness where the missing values of Y are unrelated to the value of Y itself or to any other variables in the data set. Using the income example, income is MCAR if respondents who do not report their income have the same average income as respondents who do report their income. In other words, there is no relationship at all (complete randomness) between missing values on the income variable and the values of other variables. (p. 24)

Bibliography

Chapter Four: Univariate Statistics. (n.d.). Retrieved February 8, 2020, from https://www.ssric.org/spss_manualv16/chapter4_v16.pdf

Gandrud, C. (2017). *Reproducible research with R and RStudio*.

Ho, R. (2014). *Handbook of univariate and multivariate data analysis with IBM SPSS*. CRC Press.

Horber, E. (2019). *Surveys and missing values*. http://www.unige.ch/ses/sococ/cl/stat/action/diagmiss.act.surveys.html

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Allyn and Bacon.