MISSING VALUES AND OUTLIERS IN RESEARCH DATA

JUNAID RASHID[1], KHALID WAHEED[2]
[1]*Deputy District Health Officer (DDHO), Cantt Zone, Lahore*
[2]*Department of Pulmonology & Sleep Medicine, Postgraduate Medical institute/Ameer-u-din Medical College/Lahore General Hospital Lahore.*

How to cite this article: Rashid J, Waheed K. Missing values and outliers in research data. *Pak Postgrad Med J 2020;*31(4): 167

DOI: https://doi.org/10.51642/ppmj.v31i04.404

Correspondence to: Junaid Rashid
Deputy District Health Officer (DDHO),
Cantt Zone, Lahore, Pakistan.

Email: drjunaid420@gmail.com

Missing values and outliers in data may be noticed when analyzing data for research purposes. There might be some reasons for missing values i.e., multiple people entered the data in the software, software issue or the participants did not respond to or missed some questions. Outliers may be caused by errors in data collection and software entry, equipment problems, participants answering incorrectly, or it may be a true outlier. Whatever the reason is, it becomes very frustrating for the researcher to handle this issue as it can challenge the reliability of the results[1]. Karahalios et al., reported in their review article that only 35 (43%) papers out of 82 papers included in the study, disclosed handling the missing data[2].

There are three types of missing data in research studies i.e., missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR)[3]. The cases with the missing values can be dropped out from the data as according to Schafer and Bennet dropping 5% or 10% of the data respectively does not cause biasedness in the data[4]. Another method is that missing values can be replaced (imputed) in software by series mean, mean of nearby points, median of nearby points, linear interpolation, or linear trend of point [5]. Each type of missing data has its way of handling and the researcher should take pain to handle these missing values so that the results of the study are not compromised.

Outliers are usually identified during the analysis of normality plots such as boxplots and histograms. Outliers can be managed using a trimming technique (dropping the outlier values and analyzing the rest of the data) or the winsorization technique (replacing the outlier with nearby values) [6-7]. True outliers, which cannot be overlooked, can create serious problems, and modify the study's results. Unfortunately, the literature does not help much in handling true outliers.

Journals, in my view, should provide guidelines for reporting and dealing with missing values and outliers to enable researchers to disclose this type of information. Researchers, on the other hand, must be aware of their data collection methods, correct equipment use, and data entry into software.

REFERENCES

1. Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean journal of anesthesiology. 2017;70(4):407.
2. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC medical research methodology. 2012;12(1):1-10.
3. McMahon P, Zhang T, Dwight RA. Approaches to Dealing With Missing Data in Railway Asset Management. IEEE Access. 2020;8:48177-48194.
4. Dong Y, Peng C-YJ. Principled missing data methods for researchers. SpringerPlus. 2013;2(1):1-17.
5. Cokluk O, Kayri M. The Effects of Methods of Imputation for Missing Values on the Validity and Reliability of Scales. Educational Sciences: Theory and Practice. 2011;11(1):303-309.
6. Davidov O, Jelsema CM, Peddada S. Testing for inequality constraints in singular models by trimming or winsorizing the variance matrix. Journal of the American Statistical Association. 2018;113(522): 906-918.
7. Wilcox R. Trimming and winsorization. Encyclopedia of biostatistics. 2005;8.