How to Deal with the Errors in Data Cleaning*

Bolin Shen

February 23, 2024

Data cleaning is a crucial step in the research process, shaping the foundation for statistical analyses and accurate interpretations of data findings. It requires identifying and correcting errors, inconsistencies, and inaccuracies within datasets, ensuring that the data accurately reflects the phenomenon under investigation (De Jonger 2013). In this essay, I will analyze a specific case in data cleaning. The analysis aims to expose what effects errors in data cleaning have and find out how to get rid of them in actual data analysis.

The case scenario is that the true data generating process follows a normal distribution with a mean of 1 and a standard deviation of 1. A sample of 1000 observations is collected using an instrument. The instrument's limitation (the maximum memory is 900), however, causes the last 100 observations to be duplicates of the first 100. At the same time, a research assistant unintentionally commits errors during the cleaning process, altering the sign of values and shifting decimal places (Rohan 2023). Due to the challenges posed by the errors, it is unreliable to confidently infer the true mean of the original data generating process. The following analysis of the case tries to find out the flaws, effects as well as how to address these issues.

In this case besides a flaw in the instrument used for data collection leading to the overwriting of the final100 observations with a repetition of the first 100, I also found the research assistant accidentally made two changes: First, during the data cleaning process, an error occurs as the research assistant unknowingly alters half of the negative values, changing them to positive. Additionally, the research assistant accidentally shifts the decimal place for values between 1 and 1.1. For example, a value of 1 becomes 0.1, and 1.1 becomes 0.11.

The instrumental flaw and accidental changes by the research assistant have several negative effects on the analysis. First, the redundancy caused by repeated observations may distort descriptive statistics and the underlying distribution by inflating confidence interval. The loss of unique information beyond the initial 900 observations can lead to bias in parameter estimates, affecting the accuracy of model and hypothesis testing (De Jonger 2013). Second, the unintentional changes made by research assistants during data cleansing can cause significant

*this analysis is available at: https://github.com/Brian031205/How-to-Deal-with-the-Errors-in-Data-Cleaning

distortions in the data set. These modifications can significantly affect the accuracy of descriptive statistics, including mean values and standard deviations, resulting in a distortion of the true characteristics of the original data. In addition, these changes may introduce systematic biases in the data set, potentially affecting the validity of parameter estimates and any statistical inferences (De Jonger 2013). Unintentional adjustments also raise concerns about the reliability of subsequent analyses, as altered data can influence the validity of the model, hypothesis testing, and any conclusions drawn from flawed datasets.

To ensure actual analysis having a chance to flag some of these issues during data collection and cleaning, strict quality control measures are in great need. First, establish clear documents of the instrument's limitations, such as maximum memory and overlay behavior, ensuring transparency. Regular monitoring of data patterns and ranges, especially finding duplicate sequences in the last 100 observations, can raise alerts for researchers. In addition, thoroughly checking for unexpected changes, such as changing negative values to positive values or modifying decimal points, should be part of the data clean process. When the data follow the normal distribution, the samples with a distance from the mean more than three times the standard deviation are generally considered as outliers. The specific steps are: Calculate the mean and standard deviation of the data to be tested; Compare whether each value in the data column deviates from the mean by more than 3 times, and if so, it is an outlier. By fostering a strict and careful approach to clear documentation, continuous monitoring, and validation in the data analysis, researchers can strengthen the ability of identifying and addressing such unintentional errors before they seriously impact the effectiveness of the analysis.

Reference:

Alexander, Rohan. 2023. Telling Stories with Data. https://tellingstorieswithdata.com/

De Jonge, Edwin, and Mark van der Loo. 2013. An introduction to data cleaning with R. Statistics.Netherlands:Heerlen. https://cran.r- project.org/doc/contrib/de%5FJonge+van%5Fder%5FLoo-Introduction%5Fto%5Fdata%5Fcleaning%5Fwith%5FR.pdf.

(I extend thanks to my classmates, Heng Ma, as he provided useful suggestions for this mini essay.)