The Impact of Instrumental and Processing Errors in Normal Distribution Analysis and How to Avoid them*

Yuchao Niu (Reviewed by Ziheng Zhang)

February 27, 2024

Introduction

The processes of data collection and preparation are as crucial as the process of data analysis itself. Accurate data collection and data preparation provide the foundation for robust statistical analysis. Nevertheless, different types of inaccuracies like measurement mistakes, constraints of the instruments and human errors can lead to biases and misrepresentations in the data. These inaccuracies can result in incorrect analysis and conclusions. In this paper we explored several hypothetical yet realistic situations, where errors occur during the collection and cleaning stages of a dataset. We simulated a variable with a normal distribution to explore how mistakes in data collection and cleaning can affect the analysis of such a dataset. The normal distribution, well known by its bell curve shape, is a key concept in statistics underpinning a wide array of processes in natural and social sciences. Data analyses involving variables that are normally distributed are very common therefore making this simulation very relevant to real world data analyses. R (R Core Team 2023) was the language and environment used for this essay, alongside the ggplot2 package(Wickham 2016) for data visualization.

Simulation of the True Dataset

First we performed a simulation of 1000 observations obtained from a variable that has the normal distribution and a mean of one and a standard deviation of one.

*Code and data are available at: https://github.com/MelanieNiu/Tutorial-7

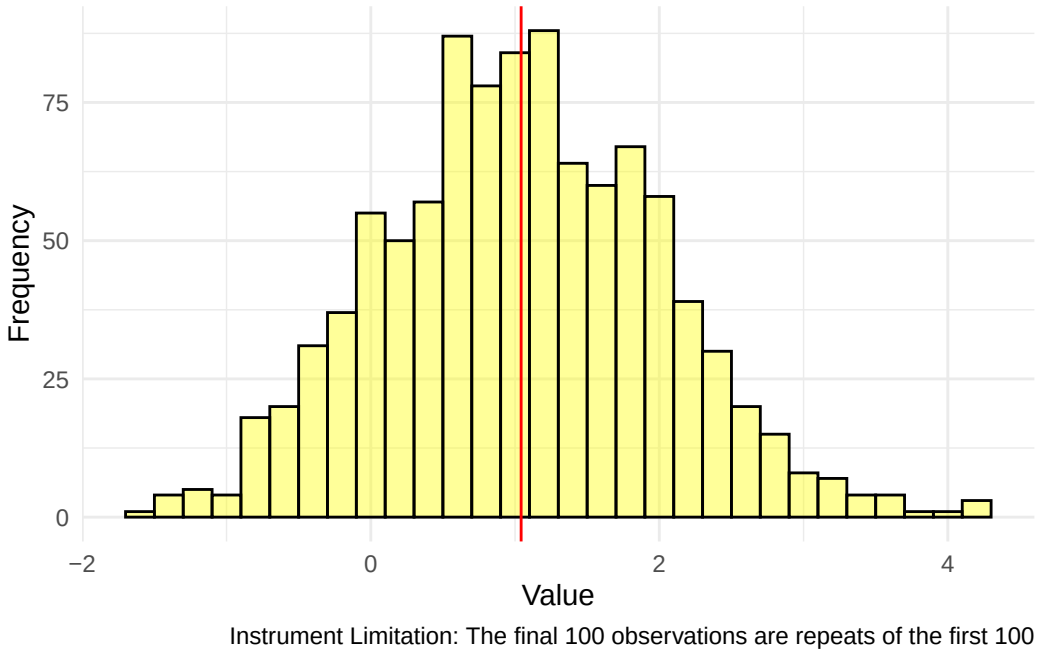

Figure 1: Data Observations are Normally Distributed

Figure 1 showed this original data of 1000 observations. The pattern of the data follows the characteristic bell curve of a normal distribution with the mean of 1.04 represented by the solid red line.

Simulation of Errors in Data Collection

Consider this situation during data collection: unknown to the data collectors, the data collection instrument has a mistake in collecting data of this sample size. The maximum memory of the instrument is recording 900 observations, and the instrument begins over-writing after 900 observations, so the final 100 observations are actually a repeat of the first 100. We simulated this instrumental error as follows.

Figure 2 demonstrated that the overwriting of the last 100 observations by the first 100 observations did not substantially alter the distribution pattern of the dataset. The new sample mean is 1.03 is close to the supposed sample mean of the dataset.

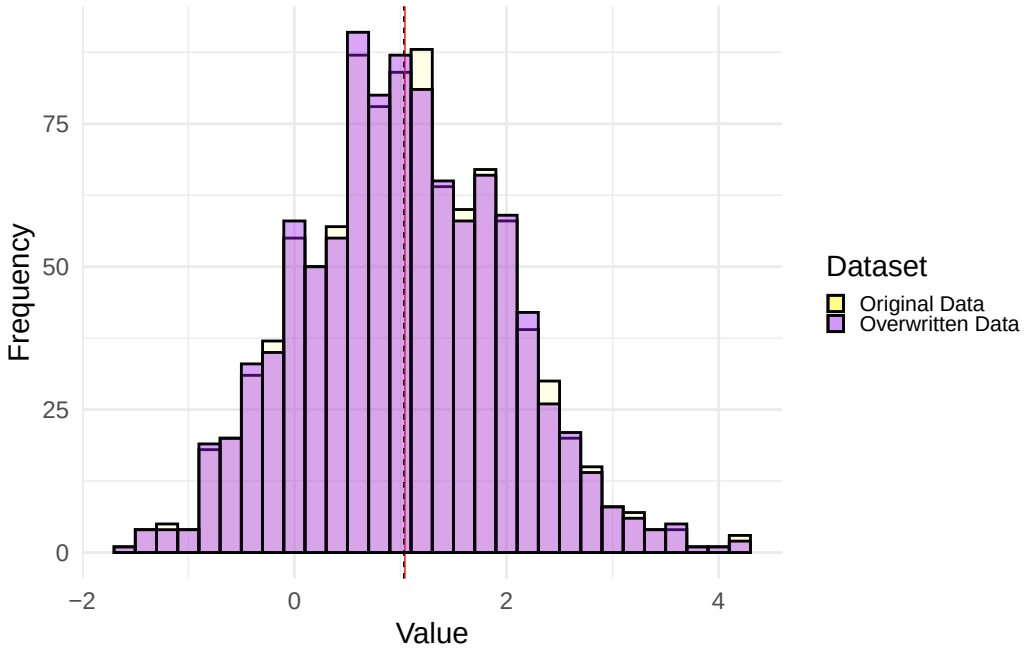

Figure 2: Impact of Instrumental Overwriting on the Dataset

Simulation of Errors in Data Cleaning

Moreover, upon the collection of data, the data set is cleaned and prepared by research assistants. During the process the research assistants accidentally changed half of the negative draws to be positive without the research team's awareness of it. We simulated the impact of this error on the data set in Figure 3.

We can see that following this accidental alteration of the dataset by research assistants, the distribution of the dataset is not substantially modified and the mean of the dataset slightly increased to 1.09. The frequency of negative values have decreased in the overwritten dataset whereas the frequency of positive values has increased as expected.

Now another mistake is made by the research assistants during data cleaning. Accidentally they change the decimal place on any value between 1 and 1.1, so that, for instance 1 becomes 0.1, and 1.1 becomes 0.11. We simulated the impact of this alteration to the dataset in Figure 4.

Figure 4 shows that following this alteration by the research assistants, There is a significant decrease in the frequency of values in the bin around 1, and an increased frequency in the values between 0 and 1. The overall distribution of the data set resembles the normal distribution less without a central peak in the bell curve shape.

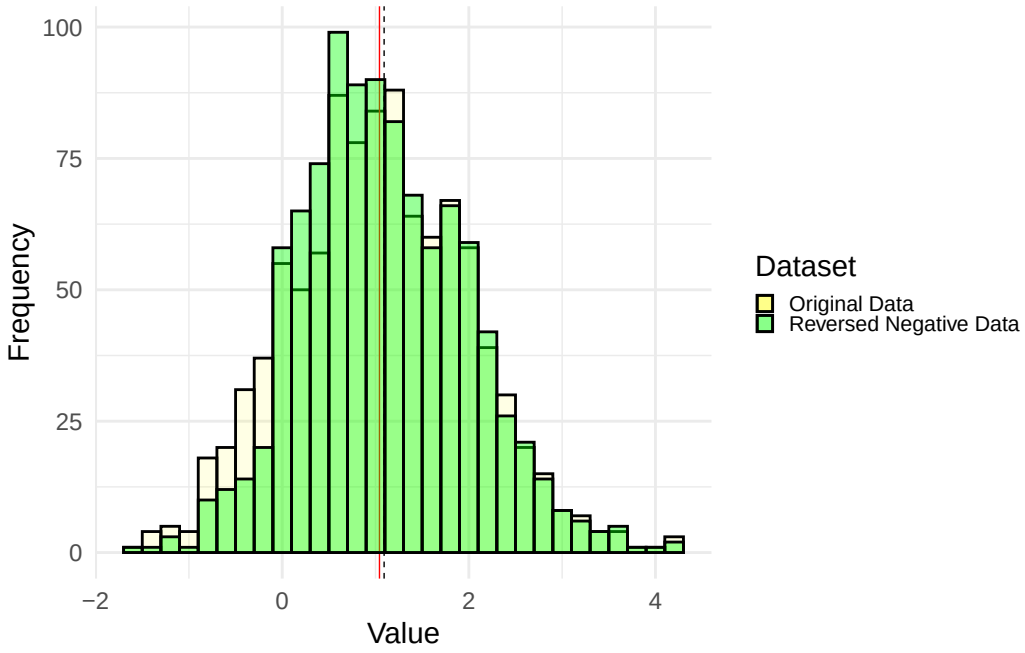

Figure 3: Impact of Reversing Half the Negative Draws to Positive

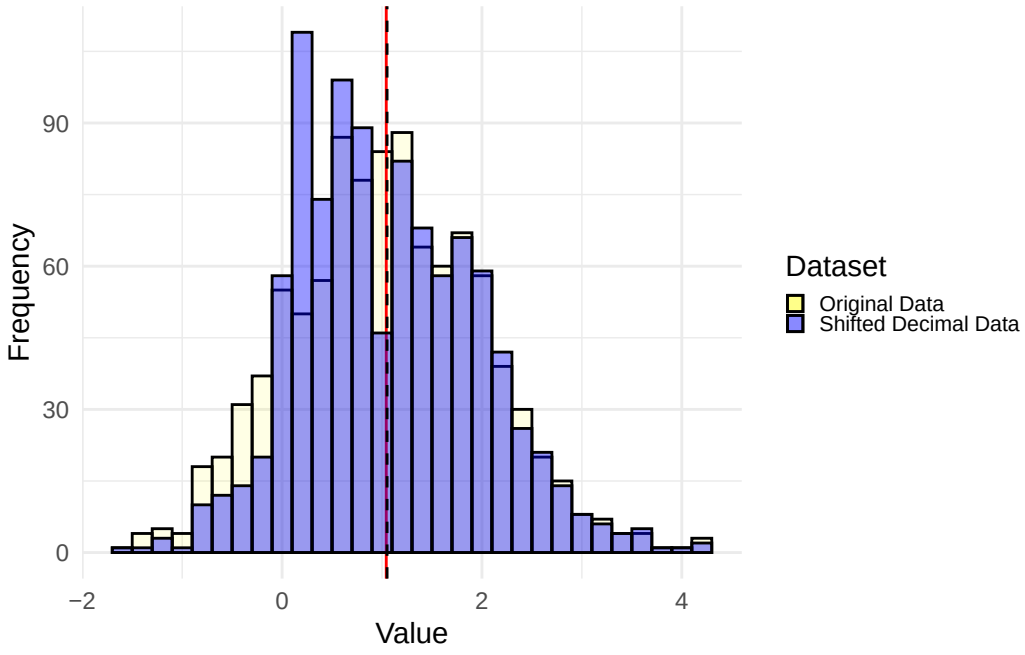

Figure 4: Impact of Shifted Decimal Places

Lastly we are interested in understanding whether the mean of the true data generating process is greater than 0 by analyzing the cleaned albeit altered dataset. We found the mean of the data set to be 1.05 which is greater than 0.

Discussions

In this hypothetical scenario, although several mistakes have happened in the analysis of a normally distributed variable with a mean of 1 and a standard deviation of 1, the final calculated mean of the dataset is still estimated to be positive and close to the true population mean. It is important to recognize that errors happen in measurement and data cleaning processes and some errors deserve priority of treatment than others. According to Broeck et al., errors which represent significant deviations within or outside the distribution of the population should be addressed with the highest priority (Van den Broeck et al. 2005). Although the errors simulated in this task did not substantially affect the sample mean, the mechanisms leading to these errors have the potential to severely compromise data integrity in real-world analyses. Therefore, it is crucial to implement data verification measures to prevent or minimize the occurrence of such errors. For instance, verification steps that can flag repetitive patterns in the dataset are useful in recognizing a chunk of data which has been overwritten by existing data. Also validation steps can be designed to recognize duplicated values in a dataset as well.

To recognize the incidences where negative draws are reversed to positive draws, we can perform checks for counts of values that fall into a particular data range. We can also apply similar checks for the counts of values between 1.0 and 1.1, as well as other data ranges, to identify mistakes made during data cleaning, as demonstrated in our hypothetical scenario.

Lastly, we can maintain detailed records of all processes undertaken during data collection, cleaning, and analysis. This approach not only aids in ensuring reproducibility but also supports data validation. By repeating the data cleaning process according to the documented procedures, we can verify whether it yields a cleaned dataset with comparable data characteristics and summary statistics.

In conclusion, data analysis can be complicated by potential and common data handling errors as discussed above even when precautions are exercised. Therefore meticulous data validation processes can guard against such errors and prove to be instrumental in ensuring data integrity. The outcome is confidence in the data and improved robustness of the analysis performed, let along a rewarding sense of accomplishment.

References

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

Van den Broeck, Jan, Solveig Argeseanu Cunningham, Roger Eeckels, and Kobus Herbst. 2005. "Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities." *PLoS Medicine* 2 (10). https://doi.org/10.1371/journal.pmed.0020267.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. https://CRAN.R-project.org/package=ggplot2.