Study of data integrity*

Zheng(Zeb) Yang

February 25, 2024

1 Methodology and Result

First of all, I set the seed to simulate the data. I began by simulating a dataset of 1,000 observations from a Normal distribution with a mean of 1 and a standard deviation of 1 and generated the histogram of the original dataset. To mimic a common instrument error, I replicated the first 100 observations at the end of the dataset, so the final 100 observations are actually a repeat of the first 100. Subsequently, I randomly changed half of the negative values to positive, simulating inadvertent data manipulation. Then, I altered the decimal place of values, change the decimal place on any value between 1 and 1.1 so that, for instance 1 becomes 0.1, and 1.1 would become 0.11, simulating a misinterpretation of data. After, upon analyzing the cleaned dataset, I generated the histogram Figure 2 of the cleaned dataset and also found that the mean of the dataset was 1.036724 and is greater than 0.

The simulation was conducted using the statistical programming language R (R Core Team 2020). To further enable the analysis I employed the use of the package of ggplot(Wickham 2016) to generate histograms.

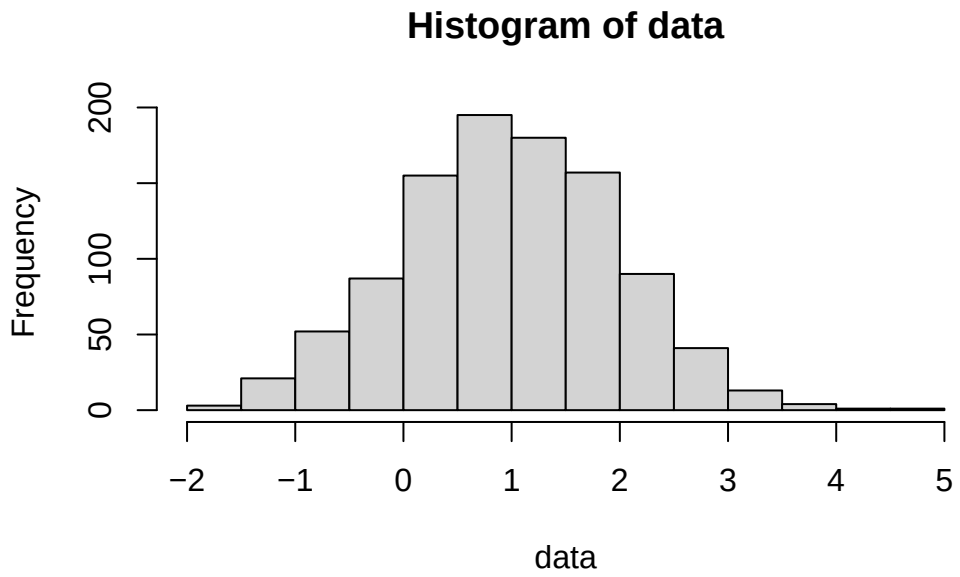

Figure 1: Histogram of the original normal distribution with mean of one and standard deviation of one

```
[1] 1.036724
```

*Code and data are available at: https://github.com/iloveyz12/Simulation-Study, Acknowledge to the review of Bernice Bao

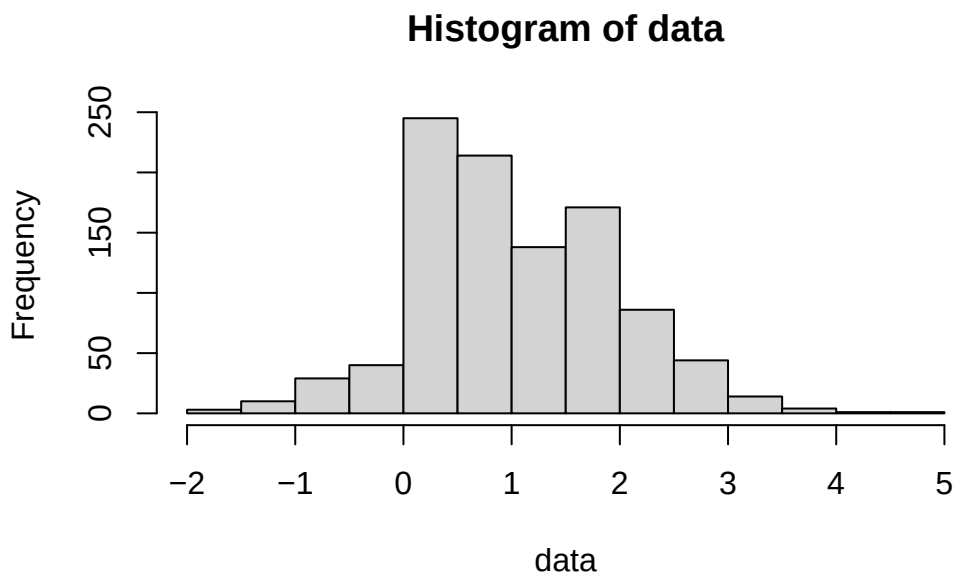

Figure 2: Histogram of cleaned dataset

2 Analysis

Figure 1 is the histogram of a normal distribution with mean of one and standard deviation of one. As you can see in Figure 1, the graph of the normal distribution is bell-shaped, with the peak of the curve occurring at the mean value, which in this case is 1. The normal distribution is symmetric, meaning that the probabilities of observing values to the left and right of the mean are equal.

However, in Figure 2, the shape of the normal distribution of cleaned dataset is notably unsymmetric and is left-skewed, which can flag the issues. The peak of the curve did not occur at the mean value one, and it occured at 0. Also, it is not symmetric at the mean value of 1. We know that the reason of the peak occuring at 0 rather than 1 is that the decimals of numbers between 1 and 1.1 had been changed, so the amount of number between 0 to 0.5 contains the numbers were originally between 1 to 1.5.

Additionally, due to the symmetry of the normal distribution, it is notable that the amount of numbers between -2 to 0 is less than the amount of numbers between 2 to 4. We know this is because the half of negative numbers were accidentally changed to positive.

The way you can falg some of the issue during an actual analysis is to apply implement automated validation to check during data collection and processing to detect anomalies, such as instrument malfunctions or data discrepancies. These checks can flag potential issues in real-time, allowing for timely intervention and correction.

References

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.