Study of data integrity[*]

Zheng(Zeb) Yang

February 25, 2024

0.1 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 1.1so 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.

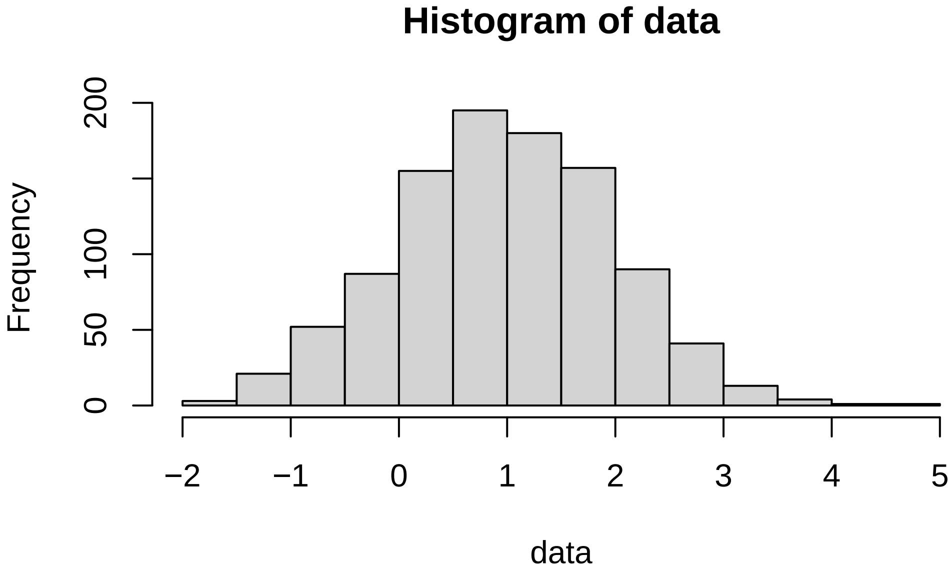

Figure 1: Histogram of the original normal distribution with mean of one and standard deviation of one

```
[1] 1.036724
```

0.2 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

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

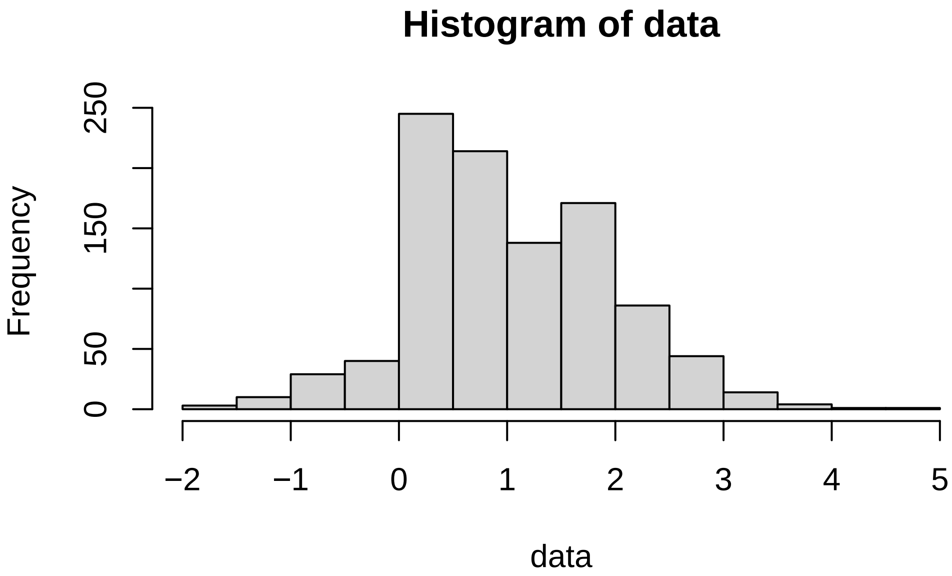

Figure 2: Histogram of cleaned dataset

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.

References

Godefroidt, Amélie. 2023. "How Terrorism Does (and Does Not) Affect Citizens' Political Attitudes: A Meta-Analysis." *American Journal of Political Science* 67 (1): 22–38. https://doi.org/10.1111/ajps.12692.

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.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.