Mini-essay 7*

Yuchen Chen

February 27, 2024

1 Introduction

In this study, we're looking to the difficulties of making sure our data is accurate in statistical analysis. We start off with a dataset that's made using a Normal distribution with a mean of 1 and a standard deviation of 1. However, the instrument has a mistake in it, which means that it has a maximum memory of 900 observations, and begins overwriting at that point. So, when we collect 1000, the last 100 get replaced with copies of the first 100. On top of that, when we're cleaning up the data, our research assistant make half of the negative numbers positive and mess up the decimal place for numbers between 1 and 1.1 by accident. Even with these mistakes, we still want to figure out if the average of the original data is more than 0 by using tool R(R Core Team (2022)). This study shows how important it is to have reliable data and talks about the problems we face with tools and human errors in data a

2 Methodology

First I generated a dataset to simulate the true data generating process, which is a Normal distribution with a mean of 1 and a standard deviation of 1. This was accomplished using a random number generator to produce 1,000 observations. Then I draw the histogram Figure 1 to see what the graph looks like normally.Next,I simulate the situation which is the measuring instrument had a memory capacity constraint that limited it to storing only the most recent 900 observations. As a result, the final 100 observations were overwritten with the first 100, thereby repeating them.Because half of the negative values in the dataset were changed to positive by the assistant, I randomly selecting half of them to be converted to their absolute values. In order to achieve this, I creates a new variable called '***negatives'*** and stores all the negative values from the'***adjusted__data'*** dataset. It does this by checking each value in

*Code and data are available at: https://github.com/Victor1114/TUT-7.gitAcknowledge the review of Simon(Pengyu Sui)

’***adjusted__data’*** to see if it is less than zero (the condition ***adjusted__data*** $< \boldsymbol{0}$), and if so, that value is included in the negatives variable. Then I use sample function to randomly select half of these negative values and use absolute function to convert them to positive equivalents. Moreover, I divide the data between 1-1.1 by 10 to simulate the described error in the data cleaning process: misplacing the decimal for numbers between 1 and 1.1.Finally, I get a cleaned dataset fits all situations, and draw a histogram Figure 2 to show the differences with initial dataset.

3 Results:

I use tidyverse(Wickham et al. (2019)) in are to estimated mean of the adjusted dataset is 1.006388. When compared to the initial dataset, which had a mean of 0.9792082, there is a slight but noticeable increase. This change in the mean is substantiated by the examination of the histograms, which illustrate the distribution of both datasets.

```
[1] 0.9792082
```

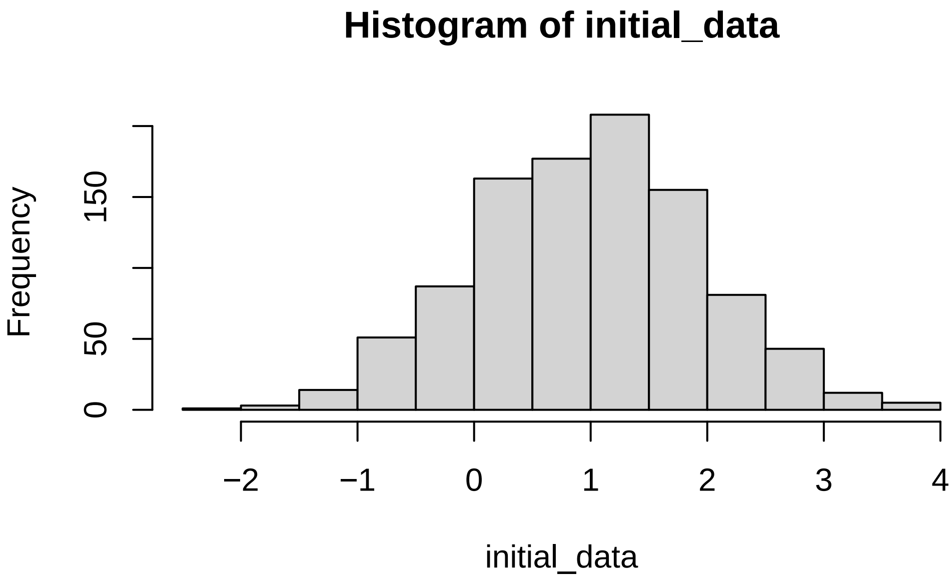

Figure 1

```
[1] 1.006388
```

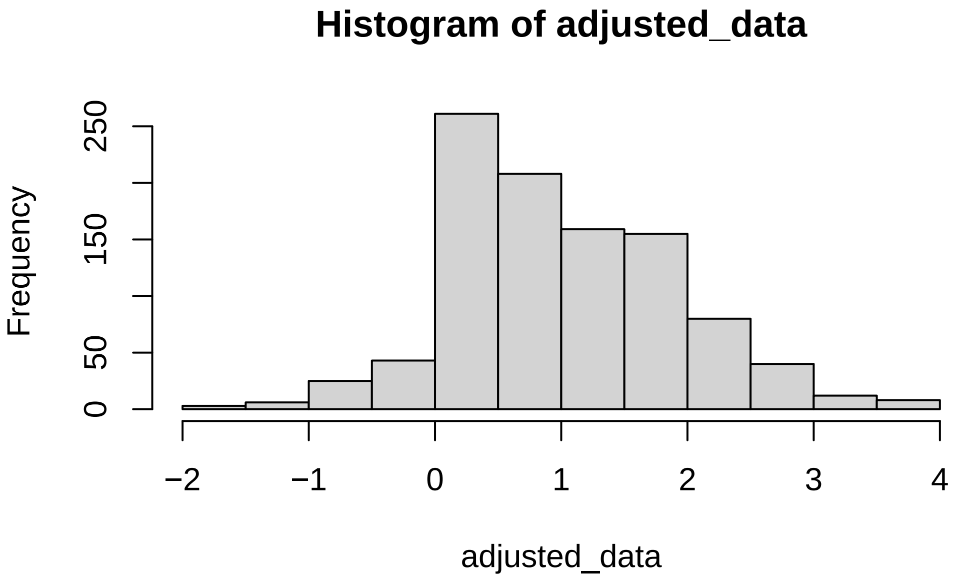

Figure 2

4 Discussion

The instrumental error resulted in the last 100 observations being duplicates of the first 100. This redundancy not only reduces the effective sample size but also biases the results if these 100 observations are not representative of the entire population.

The human error in data cleaning, where negative values were incorrectly changed to positive, results in much less negative number showing on histogram Figure 2.

The decimal place error which misplaces the decimal for numbers between 1 and 1.1 makes much more number appearing between 0-0.5 Figure 2.

5 Conclusion

The result shows that the adjusted mean is slightly higher than the mean of initial data. Even the difference not big, but it change the structure and distribution of this dataset, which should be normal distribution but it dosen't looks like. The outcomes of this research demonstrate how minor inaccuracies can make biases and emphasize the accuracy of dataset and detailed data verification to keep the reliability of statistical inferences.

Appendix

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.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.