Mini-essay 7*

Pengyu Sui

February 27, 2024

1 Introduction

In this investigation, I delve into the complexities associated with maintaining the precision of statistical data. Initially, the data is generated following a Normal distribution, characterized by a mean value of 1 and a standard deviation of 1. A limitation in the measuring instrument introduces a significant challenge; it can only store up to 900 records before it starts to overwrite existing data. Consequently, when attempting to gather 1000 entries, the final 100 entries are overwritten with duplicates of the initial 100. Compounding this issue, during the data cleansing process, an oversight leads to the alteration of negative values to positive for half of such occurrences, and the misplacement of decimal points for values ranging between 1 and 1.1. Despite these obstacles, my objective is to ascertain whether the mean of the unaltered dataset exceeds 0, utilizing the R statistical tool R Core Team (2022).

2 Methodology

To kick off this exploration, I employed a random number generator to craft a dataset that mirrors the original data creation process: a Normal distribution with a mean of 1 and a standard deviation of 1. This process yielded 1,000 individual data points. To get a visual sense of the distribution, I plotted these initial data points on a histogram, labeled as Figure 1, offering a baseline view of the dataset's distribution.

Moving forward, I addressed the challenge presented by the measuring tool's limited memory, which could only retain the latest 900 observations. This limitation led to the last 100 entries being overwritten with the first 100, effectively duplicating them within the dataset.

*Code and data are available at: https://github.com/simon0202sui/Mini-essay-7.git> Acknowledge the review of Victor Chen

The next step involved addressing the inadvertent error made during data cleaning, where negative values were mistakenly altered. To replicate this scenario accurately, I identified all negative values within the dataset, which I stored in a variable named 'negatives'. This selection was based on checking each entry in the dataset and earmarking those less than zero. With these identified, I randomly selected half of these negative values and converted them into their positive counterparts by applying the absolute value function.

Additionally, to mimic the decimal place error for numbers between 1 and 1.1, I adjusted these specific values by dividing them by 10, thereby simulating the mishap that occurred during the data cleaning process.

After implementing these adjustments, I obtained a 'cleaned' dataset that encapsulates all the discussed scenarios. To illustrate the impact of these modifications, I generated a second histogram, referred to as Figure 2. This visual comparison aims to highlight the disparities between the original and adjusted datasets, showcasing the effects of both the technological and human errors on the data.

3 Results:

In this analysis, I employed the tidyverse package Wickham et al. (2019) in R R Core Team (2022) to calculate the mean of the adjusted dataset, which resulted in a value of 1.086736. This demonstrates a modest increase when contrasted with the mean of the original dataset, which was calculated to be 1.057075. The significance of this alteration in the mean is supported through an analysis of the histograms that depict the distributions of the initial and adjusted datasets. These visual representations provide insight into how the dataset's characteristics have evolved post-adjustment, highlighting the impact of the modifications on the data's central tendency.

```
[1] 1.057075
```

```
[1] 1.086736
```

4 Discussion

The intricacies of maintaining data integrity are exemplified in this study through the examination of both instrumental and human errors in data collection and processing. The instrumental error, which led to the duplication of the first 100 observations in place of the last 100, serves as a stark reminder of the limitations inherent in data collection technologies.

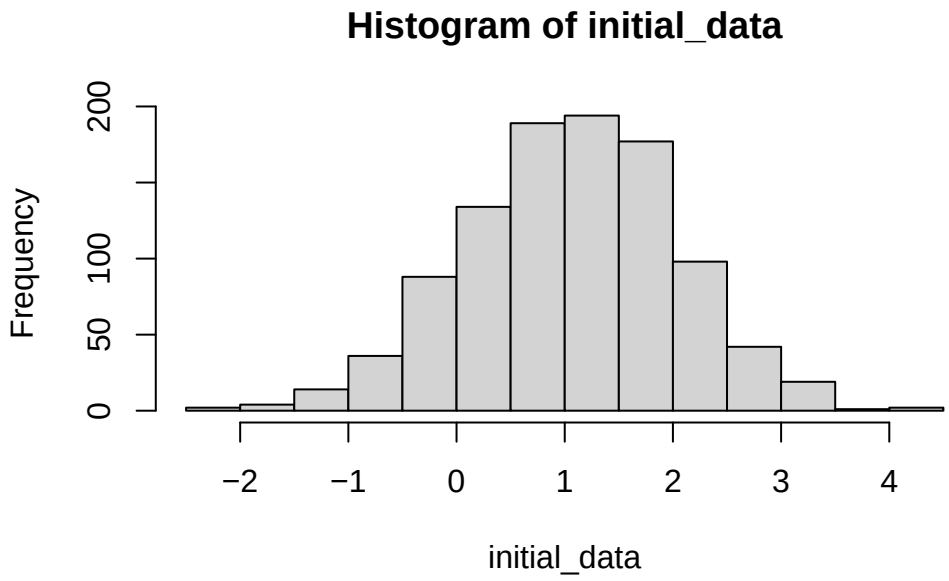

Figure 1

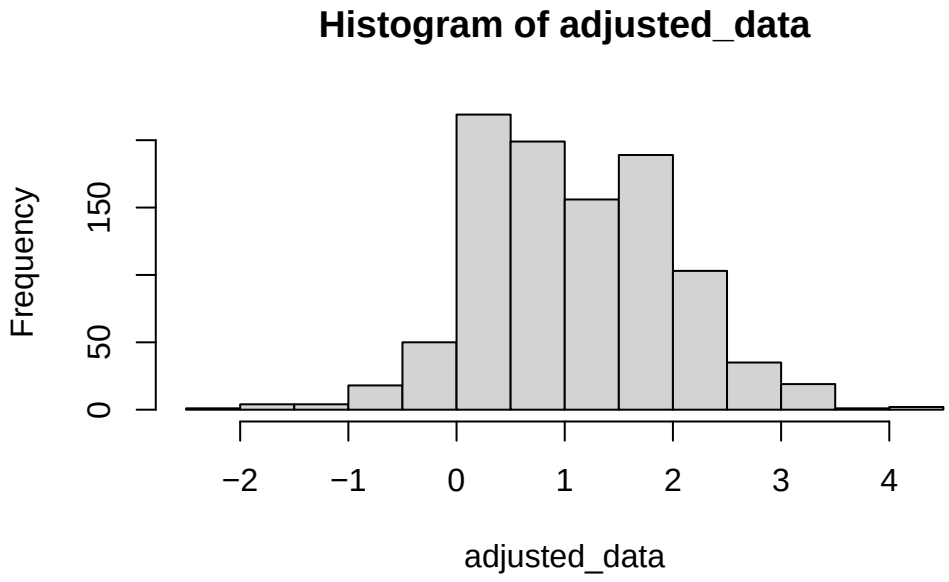

Figure 2

This redundancy not only compromises the dataset's diversity by effectively reducing its sample size but also introduces a potential bias. If these duplicated observations do not accurately represent the entire dataset, the final analysis may be skewed, offering a distorted view of the underlying statistical properties.

On the human error front, the inadvertent alteration of negative values to positive ones during the data cleaning process introduces a significant bias into the dataset. This misstep artificially decreases the number of negative observations, which is evident in the adjusted histogram (Figure 2). Such a decrease can fundamentally alter the perceived distribution of the data, impacting any conclusions drawn about the dataset's central tendencies or variability.

Moreover, the error involving the misplacement of decimal points for numbers between 1 and 1.1 further complicates the dataset's accuracy. This error results in an inflated number of observations appearing in the lower range (0-0.5) of the histogram (Figure 2). This inflation distorts the original distribution pattern, potentially leading to erroneous interpretations of the data's spread and concentration points.

These errors collectively underscore the multifaceted challenges in ensuring data accuracy. The instrumental limitation highlights the need for robust data collection tools that can prevent data loss or duplication. Simultaneously, the human-induced errors call for stringent data handling and cleaning protocols to mitigate the risk of introducing biases during data processing.

The impact of these errors extends beyond the mere alteration of mean values or distribution patterns. They compromise the dataset's integrity, potentially leading to misleading analyses and conclusions. This scenario emphasizes the importance of vigilant data management practices, including thorough validation and verification processes, to maintain the reliability and validity of statistical analyses.

5 Conclusion

The findings reveal that the mean of the adjusted dataset is marginally higher than that of the original dataset. Although this difference appears minor, it significantly impacts the dataset's structure and distribution, deviating from the expected normal distribution. This study highlights the subtle yet profound effects that even small errors can have on data integrity, altering the distribution and potentially leading to biased conclusions. It underscores the critical importance of meticulous data verification and the need for accuracy to maintain the trustworthiness of statistical analyses. Ensuring data accuracy is not merely about preserving numerical integrity but is crucial for sustaining the validity of scientific and statistical conclusions, emphasizing the role of careful data management in safeguarding against biases.

Citation

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.