Investigating the Mean of the True Data Generating Process*

Addressing Data Integrity Issues

Vanshika Vanshika

February 26, 2024

In this paper, we investigated whether the mean of a cleaned dataset, originating from a flawed instrument, exceeded zero. After analysis of data, we observed a concentration of values around zero, with a notable threshold at one. Despite data integrity issues, our findings shed light on the dataset's distribution and underscore the importance of rigorous data cleaning processes. Understanding the impact of data quality on analysis outcomes is critical for ensuring the reliability of research findings in various domains.

1 Introduction

Data integrity is crucial in statistical analysis, but various factors, including instrument limitations and human errors, can compromise this integrity. Ethical guidelines for Statistical Practice by the American Statistical Association emphasize the importance of understanding and mitigating known or suspected limitations, defects, or biases in data and methods to ensure the reliability of statistical practices (Association, 2022). This study simulates a scenario where both types of errors occur, reflecting common issues in data collection and processing. The objective is to assess the impact of these errors on statistical outcomes and propose measures to detect and prevent such discrepancies.

The remainder of this paper is structured as follows. In the Data Section, we denote how data was generated and processed. In the Results Section, we dive into the findings we discovered following the cleaned data after simulation. In the Discussion Section, we address biases and weaknesses in the data that contribute to our findings, and how we approached.

*Code and data are available at: https://github.com/vanshikav2/Linear__Model__Analysis.

2 Data

We begin by reading the cleaned dataset from the CSV file and calculating the mean of the dataset. Next, we create histograms and boxplots to visualize the distribution of the cleaned data and overlay the mean value for comparison as seen in Figure 1.

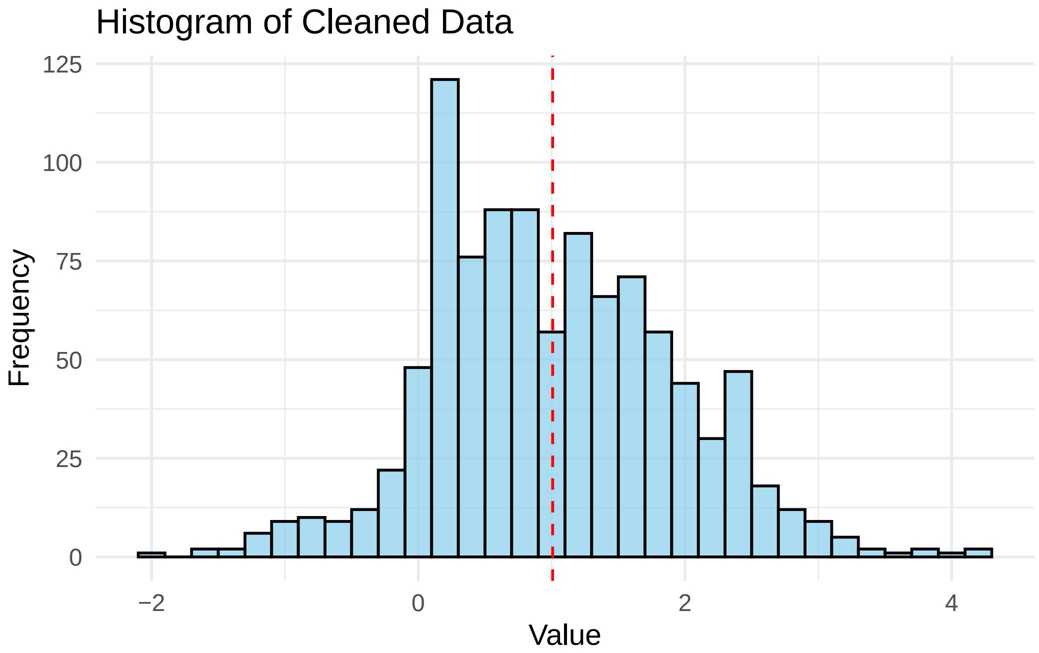

Figure 1: Mean of Cleaned Data

2.0.1 Data Cleaning

Before delving into our analysis, it is imperative to shed light on the intricacies of the data cleaning process.We used R (R Core Team 2022) for data cleaning and processing, utilizing packages like tidyverse (Wickham et al. 2019) for cleaning column names. Other packages used includes `ggplot2` (Wickham 2016), and `here` (Müller 2020). Unknown to us, the instrument used for data collection harbored a flaw, leading to the inadvertent repetition of the final 100 observations, which are duplicates of the first 100. Additionally, during the cleaning process, our research assistant inadvertently altered half of the negative values to be positive and adjusted the decimal place for values between 1 and 1.1. These inadvertent modifications necessitated thorough scrutiny to ensure the integrity of our subsequent analysis.

2.0.2 Analysis Approach

To assess whether the mean of the true data generating process is greater than 0, we employed visual analysis techniques. Specifically, we constructed a histogram of the cleaned dataset to gain insights into its distribution. Additionally, we created a boxplot to further elucidate the central tendency and variability of the dataset. These graphical representations allowed us to discern any discernible patterns, anomalies, or trends that might inform our understanding of the underlying data generating process.

3 Results

Upon examination of the histogram of the cleaned dataset, several notable observations emerged. The histogram revealed a peak frequency around 0, indicative of a substantial proportion of values clustered around this central point. This distribution pattern suggests that the dataset exhibits symmetry around 0, with values on either side being similarly represented. Notably, the presence of the red dashed line at Value = 1 on the histogram signifies a potential threshold or target value relevant to our analysis.

4 Discussion

The findings from our analysis provide valuable insights into the distribution of the cleaned dataset and its implications for evaluating the mean of the true data generating process. The observed concentration of values around 0 raises intriguing questions about the underlying data-generating mechanism and its relationship to our hypothesis regarding the mean exceeding 0. Additionally, the presence of the threshold value at 1 warrants further investigation to discern its significance in the context of our analysis.

4.0.1 Impact of Data Integrity Issues

The issues encountered during the data cleaning process had notable implications for our analysis. The instrument memory issue, resulting in the duplication of observations, introduced bias into the dataset, potentially skewing our assessment of the mean. Furthermore, the inadvertent alterations to negative values and adjustments to decimal places undermined the integrity of the dataset, necessitating careful consideration to mitigate their impact on our findings.

Steps to Ensure Data Integrity: To ensure the integrity of future analyses and preemptively flag potential issues, several proactive measures can be implemented:

1. Robust Quality Control Procedures: Establish stringent quality control protocols to detect and rectify issues during data collection and preprocessing.
2. Automated Validation Checks: Implement automated validation checks to flag anomalies and inconsistencies in the dataset, facilitating timely intervention.
3. Documentation and Transparency: Maintain comprehensive documentation of data collection and cleaning procedures, enhancing transparency and reproducibility.
4. Peer Review Processes: Incorporate peer review mechanisms to validate findings and identify potential sources of bias or error, fostering rigor and accountability in the analysis process.

By adopting these measures, researchers can bolster the reliability and validity of their analyses, thereby enhancing confidence in the resulting findings and facilitating meaningful contributions to the body of knowledge in their respective domains.

References

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

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. 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.