Clean and prepare of data*

How mistake affect the conclusion

Zixuan Yang

February 27, 2024

Data cleaning and preparation are crucial steps in research, but they can inadvertently introduce systematic biases that affect the accuracy of subsequent analyses. This paper investigates the impact of such biases on data distribution and hypothesis testing through simulated scenarios. Using a simulated dataset with known errors, we demonstrate how these biases can lead to erroneous conclusions, highlighting the importance of vigilant data management practices in research. Understanding and addressing systematic biases are essential for ensuring the reliability and integrity of research findings.

1 Introduction

Data obtained through surveys or experiments often come with various issues such as missing data and anomalies, which can affect subsequent research analysis. Therefore, after acquiring raw data, it is essential to undergo data cleaning and preparation to obtain the final dataset. However, during the process of data cleaning and preparation, there may be operational errors that introduce systematic biases Chu et al. (2016). These systematic biases can significantly impact the accuracy of subsequent analyses and cannot be resolved through further analysis, posing a fatal error to research conclusions.

Systematic biases may exist in various aspects, including data overflow, symbol errors, and decimal point entry errors. This paper aims to investigate the effects of these biases on data distribution and hypothesis test conclusions through simulating systematic bias methods.

The remainder of this paper is structured as follows. Section 2 will introduce introduce the simulation of the original data and the errors introduced during the data cleaning process. Section 3 will present the distributions of the original data and the erroneous data, and we will employ hypothesis testing to demonstrate the erroneous conclusions drawn from the error

*Code and data are available at: LINK.

data. Section 4 will delve into the discussion on how to address these systematic biases and enhance the quality of our research.

2 Data

Assume the true data generating process is a Normal distribution with mean of one, and standard deviation of 1. We obtain a sample of 1,000 observations using some instrument using R Core Team (2022). However, there are some mistakes happened during the data prepare.

1. Unknown to us, the instrument has a mistake in it, which means that it has a maximum memory of 900 observations, and begins over-writing at that point, so the final 100 observations are actually a repeat of the first 100.
2. We employ a research assistant to clean and prepare the dataset. During the process of doing this, unknown to us, they accidentally change half of the negative draws to be positive.
3. They additionally, accidentally, 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.

After these simulations, I will use the processed data to do the hypothesis test in the next section.

3 Results

Figure 1 depicts the distribution of original and processed data using a histogram plot generated with ggplot2 by Wickham (2016). The two plots are combined by the package Auguie (2017). It is evident that after processing, the data is skewed towards the range of 0 to 1, with a departure from the typical bell-shaped curve. Furthermore, the results of the one-sided t-test indicate a test statistic of 12.612 and a p-value approaching 0. Based on this hypothesis test, there is compelling evidence to reject the null hypothesis and conclude that the mean of the data is greater than 0. However, this outcome contradicts the nature of the original data, suggesting that errors introduced during data manipulation can render research conclusions invalid.

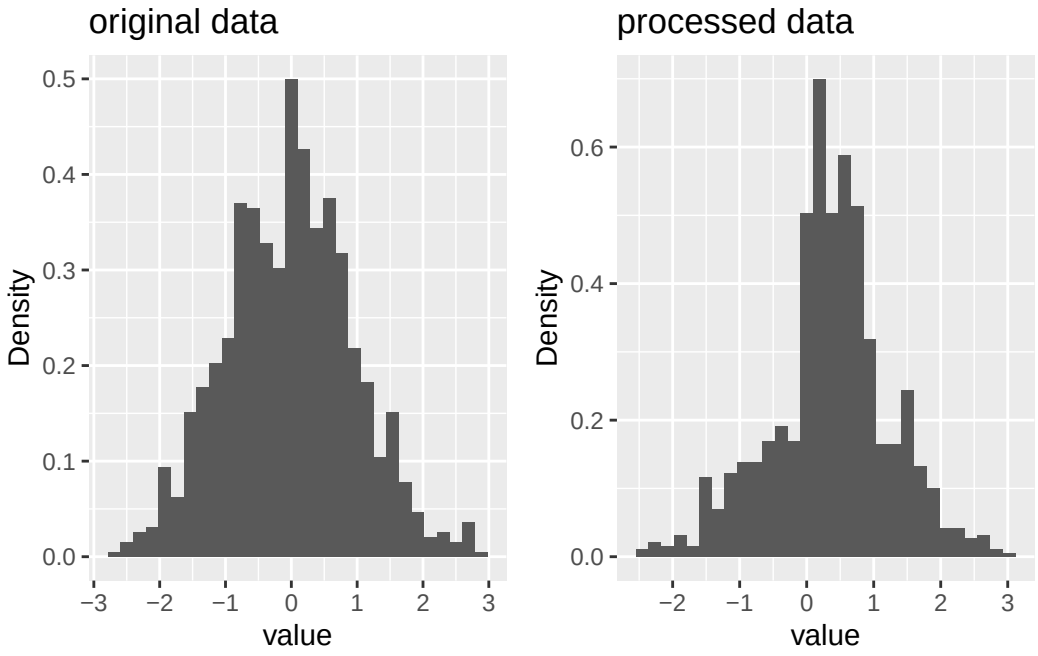

Figure 1: Comparison of orignal and processed data

4 Discussion

Despite the importance of data cleaning and preparation, it is essential to acknowledge the potential introduction of systematic biases during these processes. Systematic biases can arise due to human error, measurement errors, or systematic differences in data collection methods. These biases can skew the results of analyses and lead to incorrect conclusions if not properly addressed.

Therefore, researchers must be vigilant in identifying and mitigating systematic biases during data cleaning and preparation. This may involve conducting sensitivity analyses to assess the robustness of results to different cleaning and preparation methods or implementing quality control procedures to minimize the risk of bias introduction. By understanding and addressing systematic biases, researchers can ensure the integrity and reliability of their data analyses and research findings.

References

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Chu, Xu, Ihab F Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. "Data Cleaning: Overview and Emerging Challenges." In *Proceedings of the 2016 International Conference on Management of Data*, 2201–6.

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.