Navigating the Pitfalls of Data Cleaning: Avoiding Incorrect Conclusions*

Understanding the Impact of Mistakes on Research Outcomes

Chenhang Huang

February 27, 2024

This study investigates the impact of systematic biases introduced during the data cleaning process on research outcomes. We simulate common mistakes such as memory issues, data entry errors, and decimal shifts to assess their effects. The results highlight significant differences between the original and cleaned data, emphasizing the importance of thorough data validation procedures. Addressing these biases is crucial for ensuring the reliability and validity of research conclusions.

1 Introduction

After we get the original data, we have to clean and prepare it to make the final dataset. But sometimes mistakes happen during this process, causing problems called systematic biases. These biases can mess up our analysis and lead to wrong conclusions that we can't fix later. They can happen in different ways, like getting too much data or making mistakes with symbols or decimals. This study looks at how these mistakes affect the data and our conclusions, by simulating them. We want to learn more about how these mistakes can affect our research and why it's important to be careful when cleaning data.

The remainder of this paper is structured as follows. in Section 2, we present the simulation of the data processing procedure.Then, Section 3 compares the difference between original data and the cleaned data. Section 4 summarize the findings and discuss the consequences of errors in data cleaning.

[*] Code and data are available at: LINK.

2 Data

In this section, I'll begin by sampling 1000 units from a normal distribution with mean 0 and standard deviation 1 to represent the true values. Next, I'll simulate three mistakes in the data cleaning process using the method described in (R Core Team 2022). These mistakes include: 1) the instrument has a memory issue causing the last 100 observations to be a repetition of the first 100; 2) accidentally changing half of the negative values to positive during data cleaning; 3) accidentally shifting the decimal place for values between 1 and 1.1. After cleaning, I'll compare the true data with the cleaned data to highlight any differences.

Table 1: Summarize of the original and new data

x	newx
Min. :-3.008049	Min. :-2.93977
1st Qu.:-0.656863	1st Qu.:-0.03711
Median :-0.036163	Median : 0.41708
Mean :-0.001166	Mean : 0.38502
3rd Qu.: 0.689635	3rd Qu.: 0.95987
Max. : 3.810277	Max. : 3.81028

3 Results

Our results are summarized in Table 1 using the package knitr by Xie (2015). While the mean and median of the original data are close to 0, those of the new data are 0.42 and 0.39 respectively, indicating notable differences between the two datasets. I also show their distributions in Figure 1 by using Wickham (2016). It can be seen that the original value is near normal distribution, but the new data shape is not bell shape. In the new data, most of the values are within 0 and 1. I also run the hypothesis test for the new dataset, the null hypothesis is :

$$H_0 : \mu = 0$$

$$H_1 : \mu > 0$$

the t statistic is 12.88, which is higher than 2, so we reject the null hypothesis. For the original data, the t-statistic is only -0.036, near zero, and we fail to reject the null hypothesis.

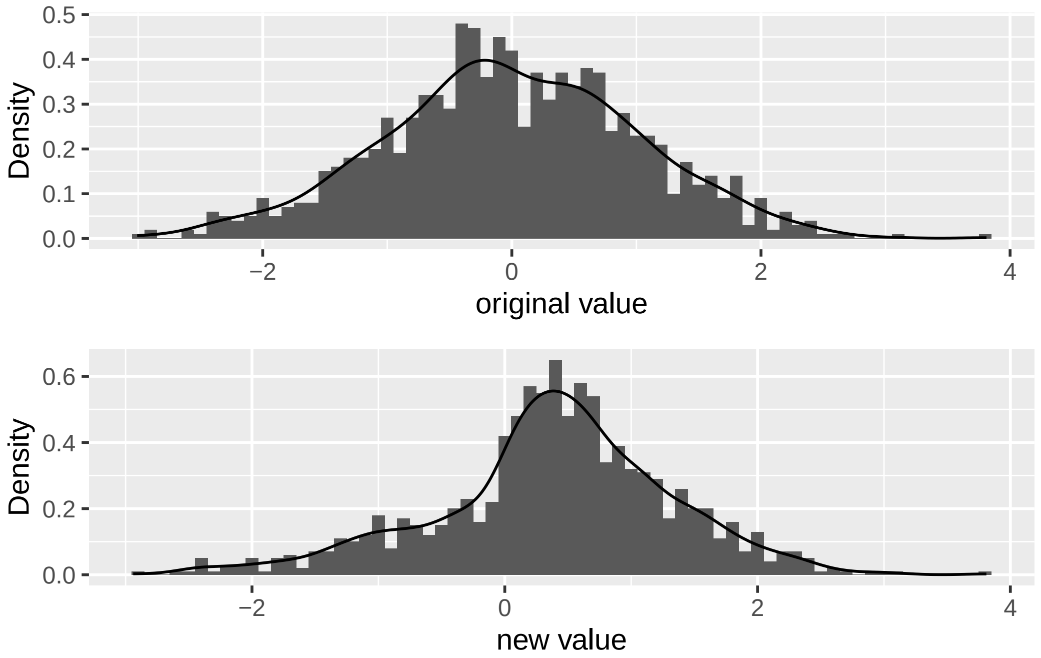

Figure 1: Distribution of original and new data

4 Discussion

Through the above analysis, we find that errors introduced during data processing can lead to significant discrepancies between the processed and original data. The biases introduced in the simulation are primarily due to errors in instrument measurement and data entry. This outcome serves as a reminder that it's crucial to check the accuracy of data during processing and conduct effective data testing to avoid the generation of such systematic errors.

References

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.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.