Understanding the Impact of Data Preprocessing Errors on Statistical Analysis*

A Case Study

Chay Park

In this paper, we investigate the effect of data preprocessing errors on statistical analysis using a simulated dataset. We simulate a situation where the data generating process follows a Normal distribution with a mean of one and a standard deviation of one. However, due to errors in the data collection and preprocessing stages, the dataset undergoes several transformations. Specifically, we introduce three types of errors: (1) data truncation due to limited instrument memory, (2) alteration of negative values to positive values, and (3) incorrect adjustment of decimal places. Subsequently, we analyze the impact of these errors on the mean estimation and discuss potential steps to identify and mitigate such issues in real-world data analysis.

1 Understanding the Impact of Data Cleaning Errors on Statistical Analysis

1.1 Introduction

Accurate data preprocessing is crucial for obtaining reliable results in statistical analysis. However, errors in data collection and preprocessing can introduce bias and distort the underlying data distribution. In this paper, we examine the effects of data preprocessing errors on statistical inference using a simulated dataset. We simulate a scenario where the true data generating process follows a Normal distribution with a known mean and standard deviation. Subsequently, we introduce errors during the data preprocessing stage to mimic real-world data issues. We aim to assess the impact of these errors on estimating the mean of the true data generating process and discuss strategies to mitigate such errors in practice.

*Code and data are available at: https://github.com/Chay-HyunminPark/A-Case-Study.git

1.2 Data Simulation and Preprocessing Errors

We begin by simulating a dataset of 1,000 observations from a Normal distribution with a mean of one and a standard deviation of one. However, we introduce three types of errors during the data preprocessing stage. Firstly, we impose a limitation on the instrument's memory, causing it to overwrite the final 100 observations with a repeat of the first 100 observations. Secondly, we instruct the research assistant to randomly change half of the negative values to positive values, introducing additional noise into the dataset. Lastly, we direct the research assistant to incorrectly adjust the decimal places for values between 1 and 1.1, leading to further distortions in the dataset.

1.3 Effect of Errors on Statistical Analysis

Next, we analyze the impact of these errors on estimating the mean of the true data generating process. Due to the truncation error, the repeated observations at the end of the dataset skew the distribution, potentially inflating the estimated mean. The alteration of negative values to positive values introduces noise and biases the mean estimate. Additionally, the incorrect adjustment of decimal places distorts the distribution of values between 1 and 1.1, further affecting the mean estimation. Overall, these errors collectively lead to an inaccurate estimation of the true mean.

1.4 Inferential Analysis

Next, we conduct inferential analysis to determine whether the mean of the true data generating process is greater than 0. We formulate hypotheses and perform hypothesis testing, comparing the sample mean to the hypothesized population mean of 0. However, it is crucial to consider the potential biases introduced by the data cleaning errors and their impact on the validity of our statistical inference.

1.5 Mitigation Strategies

To mitigate the impact of data preprocessing errors on statistical analysis, several strategies can be implemented. Firstly, thorough data validation and verification procedures should be conducted during the data collection and preprocessing stages to identify and rectify any errors promptly. Secondly, employing robust statistical techniques that are less sensitive to outliers and data distortions can help mitigate the effects of preprocessing errors on inference. Additionally, implementing automated data quality checks and validation scripts can help flag potential errors and anomalies in the dataset before analysis.

1.6 Conclusion

In conclusion, this paper highlights the importance of accurate data preprocessing in statistical analysis and demonstrates the potential impact of preprocessing errors on mean estimation using a simulated dataset. The errors introduced during data collection and preprocessing stages can significantly distort the underlying data distribution and lead to biased inference. By understanding the effects of these errors and implementing appropriate mitigation strategies, researchers can enhance the reliability and validity of their statistical analysis in real-world scenarios.

2 References