Impact of Data Errors on Statisitcal Analyses*

Shivank Goel

February 23, 2024

This paper explores the consequences and impact of instrument and human errors in data collection and processing, using a simulated dataset. The actual process that produces the data, assumed to follow a Normal distribution with a mean of one and a standard deviation of one. This data was first simulated, however, the instrument used for data collection had an issue. It resulted in last 100 observations being overwritten by the first 100. Additionally, during data cleaning, negative values were mistakenly changed to positive, and decimal places were incorrectly moved for values between 1 and 1.1. The study examines the impact of these errors on the dataset's mean and discuss strategies to detect and address such kind of mistakes in real-world scenarios.

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 American Statistical Association states: "Integrity of Data and Methods: The ethical statistical practitioner seeks to understand and mitigate known or suspected limitations, defects, or biases in the data or methods and communicates potential impacts on the interpretation, conclusions, recommendations, decisions, or other results 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.

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

*Code and data are available at: https://github.com/shivankgoel003/Simulation-of-Data-with-Instrument-and-Human-Errors

the data that contribute to our findings, and how we approached.The last section is Conclusion and Acknowledgements.

2 Data

The data was generated to represent a Normal distribution (mean = 1, SD = 1) with 1,000 observations. Due to instrument limitations, the last 100 observations were overwritten by the first 100. During data cleaning, half of the negative values were altered to positive, and values between 1 and 1.1 had their decimal places shifted. The R programming language (R Core Team 2022) was used for both data simulation and cleaning processes. A sample of cleaned state score data can be seen in Table 1.

Table 1: Sample of Cleaned Data

Observation
0.4395244
0.7698225
2.5587083
0.1070508
1.1292877
2.7150650

3 Results

The cleaned data exhibited a mean value slightly deviating from the true mean of the original distribution. This deviation can be due to overwriting and data cleaning steps. The overwriting error created a repetition in the dataset, resulting in reducing its variability and leading to biased estimates. The cleaning errors further changed the data, particularly the decimal shift, which significantly altered the values within a specific range.

4 Discussion

The study highlights the possibility of measurement errors in statistical analysis in data collection and processing. Instrument limitations, such as memory constraints, can lead to significant data loss or other issues, as seen in the overwriting error. Human errors during data cleaning, often due to oversight, can further lead to errors in data processing.

Therefore, in order to reduce these errors, it is essential to perform checks for instrument and other possible flaws in data. A validation or testing step in the data cleaning process, can help identify and correct errors. Furthermore, training and clear documentation for data handling are crucial.

5 Conclusion and Acknowledgements

This study highlights the importance of integrity in data collection and processing. Understanding the nature and impact of errors can improve the statisitical analyses and ensure more accurate findings.

Special thanks to Vanshika Vanshika for peer review and providing essential feedback and suggestions.

References

Association, American Statistical. 2022. *Ethical Guidelines for Statistical Practice*. https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.