Comparing Falsely Processed Data with True Underlying Distribution*

Yang Zhou

February 27, 2024

Table of contents

1 Introduction

In this study, we investigate the implications of data manipulation errors on statistical analysis outcomes. We simulate a scenario where the data generation process is assumed to follow a Normal distribution with a mean of one and a standard deviation of one, generating a sample of 1,000 observations. This simulation incorporates three specific data manipulation errors: overwriting of observations due to instrument memory limitations, accidental sign changes of negative values, and decimal place errors. Our goal is to assess the impact of these errors on the mean estimation and to explore methodologies that could detect such discrepancies in real-world data analysis scenarios.

*Code and data are available at: https://github.com/yangzhoucoco/omparing-Falsely-Processed-Data-with-True-Underlying-Distribution

2 Data Simulation

The data was generated and manipulated using R programming (R Core Team 2022), and using ggplot2 (Wickham 2016) for visualization. The simulation process involved the following steps:

1. Initial Sample Generation: A sample of 900 observations was drawn from a Normal distribution with a mean of 1 and a standard deviation of 1.
2. Observation Overwriting: Due to instrument memory limitations, the final 100 observations were a duplication of the first 100, simulating an error in the data recording process.
3. Sign Correction Error: Half of the negative values in the sample were mistakenly converted to positive values, introducing a bias.
4. Decimal Place Error: Values between 1 and 1.1 had their decimal places inaccurately shifted, altering their magnitude erroneously.

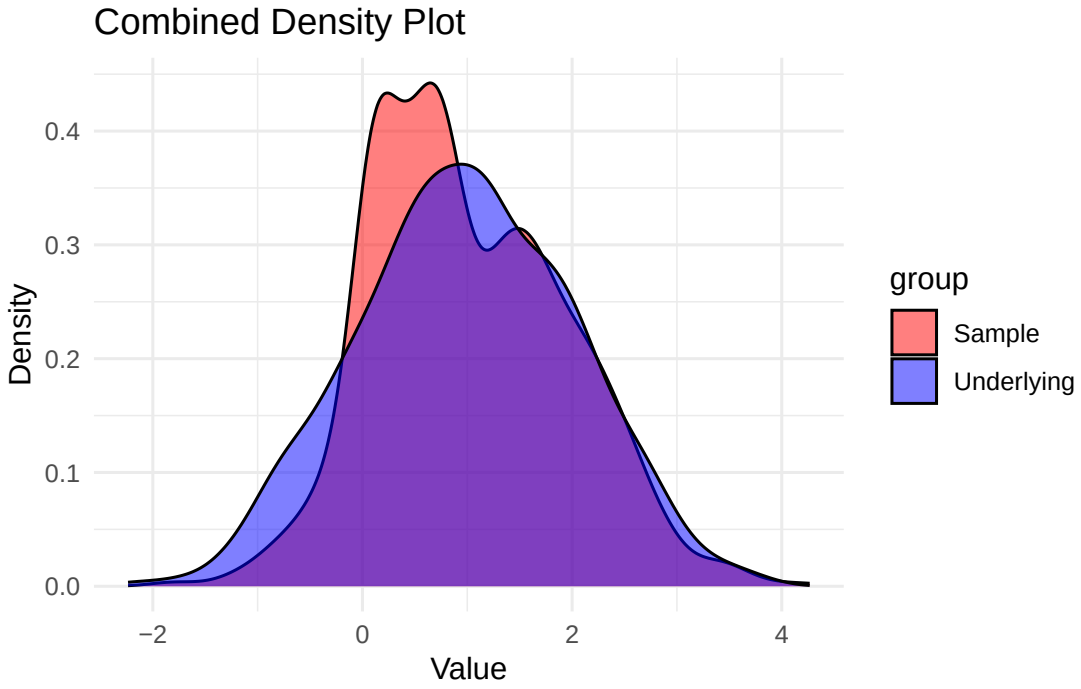

Figure 1: Density graphs of simulated data and true distribution

3 Results

The density plots Figure 1 revealed significant deviations between the manipulated sample and the underlying true distribution. The manipulations introduced a visible bias towards positive values and altered the shape of the distribution, particularly around the mean and the left tails.

4 Mitigation Strategies

To prevent such errors from compromising real-world data analysis, several steps can be recommended:

1. Data Integrity Checks: Implementing checks to verify the integrity of the data, including maximum memory checks and validation of value ranges.
2. Frequent Communication with Data Handlers: Ensuring clear communication channels with individuals involved in data collection and processing to quickly identify any unintended manipulations.

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.