# Impact of Data Processing Errors on Statistical Analysis

Pu Yuan

2024/2/25

## Introduction

In this analysis, we explored a dataset that was generated to simulate a series of errors that might occur in real-world data collection and processing. The objective was to understand the impact of these errors on statistical analyses, specifically on estimating the mean of the underlying data generating process. The data was initially generated to follow a normal distribution with a mean of 1 and a standard deviation of 1. However, due to instrument limitations and processing errors, the final dataset underwent significant alterations. These alterations included the overwriting of the last 100 observations due to memory limitations, changing half of the negative values to positive, and misadjusting the decimal places for values between 1 and 1.1.

This paper uses R(R Core Team 2023a) and R package states(R Core Team 2023b), and has been updated based on the feedback of Yiyi Yao.

## Simulated Errors and Their Impacts

Instrument Memory Limitation: The first 100 observations were repeated as the last 100 due to the instrument's memory limitation. This repetition artificially increased the sample's homogeneity, potentially biasing any analysis towards the characteristics of these 100 observations.

Changing Negative Draws: By converting half of the negative observations to positive, the variability in the data was reduced, and the mean was artificially inflated. This action skewed the distribution of the data and misrepresented the underlying variability present in the original dataset.

Decimal Place Adjustment: The misadjustment of decimal places for specific values led to a significant distortion of the data scale for a subset of observations. This error could severely impact the mean and variance estimates, leading to incorrect inferences about the data.

Finally, we got the cleaned data and have 95% confidence level that the mean is greater than 0. The estimated mean value is 0.9921027.

```
## [1] 0.9921027
```

```
##
##  One Sample t-test
##
## data:  original_data
## t = 31.477, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.9402123       Inf
## sample estimates:
## mean of x
## 0.9921027
```

# Mitigation Methods

Perform Initial Checks: Before analysis, assess the data for basic quality indicators such as the range, distribution, and presence of expected patterns or values. This can help identify obvious issues such as duplicated data or unrealistic values.

Check for Duplicates: Specifically look for duplicated records, which in this case could help identify the overwritten data points.

Check Data Length: Ensure the dataset length matches the expected number of observations. A mismatch may indicate overwriting or data loss.

# References

R Core Team. 2023a. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

———. 2023b. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.