Simulating and Analyzing Instrument and Human Error in Statistical Data Analysis*

Yingqi Pang

27 February 2024

Contents

Introduction

The process of data analysis is critical to extracting meaningful insights from information. However, the integrity of data is paramount; any error in data collection or processing can lead to inaccurate conclusions. In this study, we simulate a scenario involving both instrument and human errors to assess their impact on the analysis of a normally distributed dataset.

Methodology

We began with the generation of a dataset presumed to be from a normal distribution with a mean (μ) of 1 and a standard deviation (σ) of 1. A sample of 1,000 observations was generated using R's random normal distribution function `rnorm()`. This dataset represents an ideal state without any errors.

We then introduced two types of errors:

1. **Instrument Error**: A memory limitation error was simulated where the last 100 observations of our dataset were overwritten with the first 100 observations. This error represents a common issue in data collection where instrument limitations lead to data loss or corruption.
2. **Data Cleaning Error**: During the data cleaning process, two mistakes were simulated:
    - Conversion of half of the negative values to their positive counterparts, thus altering the distribution of the data.
    - Adjustment of the decimal place for values between 1 and 1.1, incorrectly reducing these values by an order of magnitude (e.g., 1.05 became 0.105).

The dataset, now containing both instrument and human errors, was then analyzed to estimate the mean and to perform a one-sided t-test. The t-test was used to test the hypothesis that the mean of the data generating process is greater than zero.

*Code and data are available at: https://github.com/pangyin2/-Simulating-and-Analyzing-Instrument-and-Human-Error-in-Statistical-Data-Analysis.git

Findings

Upon analyzing the 'cleaned' dataset:

- The **mean** was calculated, which reflects the central tendency of the dataset after the introduction of errors.
- A **one-sided t-test** was conducted to determine if the mean of the dataset was significantly greater than zero.

The analysis did not proceed without the recognition that the errors introduced would likely distort the findings:

- The **instrument error** would introduce a significant bias as the latest data points were not new observations but repeats. This would likely inflate the mean if the initial observations had a higher mean than the true distribution.
- The **data cleaning error** that changed negative values to positive would artificially increase the mean, as negative deviations from the true mean would be underrepresented.
- The **decimal place error** would disproportionately affect a subset of the data (values between 1 and 1.1), leading to a lower mean than expected, although this impact might be relatively small given the specific range of values affected.

Discussion

The simulated errors had a pronounced effect on the data analysis, potentially leading to erroneous conclusions. The mean calculated from the erroneous data would not be a true reflection of the underlying normal distribution. The t-test result would be questionable as the data no longer followed the assumed normal distribution due to the introduced biases.

Steps for Mitigation

To avoid such errors affecting the analysis, several steps can be implemented:

1. **Instrument Validation**: Regular checks and calibrations should be performed to ensure that all instruments are functioning correctly and within their limits.
2. **Data Integrity Checks**: Implement automated checks to identify unusual patterns in the data that could indicate overwriting or duplication.
3. **Cleaning Validation**: Data cleaning processes should be standardized and include validation steps. Any transformations made to the data should be documented and reversible.
4. **Outlier Analysis**: Conducting outlier analysis could help in identifying and investigating anomalies that could be due to data errors.
5. **Audit Trails**: Maintaining audit trails of both the original and cleaned data sets along with the cleaning procedures applied would allow for traceability and verification of the data cleaning process.
6. **Sensitivity Analysis**: Performing sensitivity analyses to understand how robust the findings are to different assumptions about the data can help in assessing the impact of potential errors.

Conclusion

The simulation study underscores the profound impact that instrument and human errors can have on data analysis. It highlights the need for rigorous data validation and cleaning processes to ensure the accuracy of statistical analyses. By implementing the aforementioned steps, the likelihood of such errors leading to incorrect conclusions can be significantly reduced.