Identifying and reduce Data Collection and Cleaning Errors in Analysis*

Kuiyao Qiao

April 3, 2024

This report demonstrates how data collection and cleaning errors can bias statistical analyses and lead to incorrect conclusions. Analysis of the resulting data shows that they deviate significantly from the true parameters, and inferences are incorrect. The implications of each error are discussed, and strategies for identifying and mitigating data quality problems are proposed, emphasizing the importance of scrutiny throughout the analysis.

1 Introduction

In statistical analysis, the validity of our conclusions is fundamentally dependent on the quality of the data we use. Even with a well-designed study and appropriate analytical methods, undetected data collection and cleaning errors can introduce substantial bias and lead to incorrect inferences. R Core Team (2023)

2 Simulation Setup

The actual data-generating process in this scenario is a Normal distribution with a mean of 1 and a standard deviation of 1. Using some instruments, we aim to obtain a sample of 1,000 observations from this distribution. However, unknown to us, several errors occur during data collection and cleaning:

1. The instrument has a maximum memory of 900 observations and begins overwriting old values after this point. As a result, the final 100 observations are a duplicate of the first 100.

*Code and data are available at: https://github.com/QPP123/Identifying-and-reduce-Data-Collection-and-Cleaning-Errors-in-Analysis

2. During data cleaning, a research assistant accidentally converts half of the negative values in the dataset to positive.

3. Additionally, for any values between 1 and 1.1, the research assistant unintentionally shifts the decimal place to the left by one digit (e.g. 1 becomes 0.1, 1.05 becomes 0.105).

After receiving this "cleaned" dataset, I want to determine whether the mean of the true data-generating process is greater than 0.

3 Simulation and Analysis

First 1,000 draws were simulated from a Normal(1,1) distribution in R to represent the true data-generating process. Then, the data collection error was reproduced by replacing the last 100 values with a duplicate of the first 100. Next, I randomly selected 50% of the negative values, converted them to positive, and shifted the decimal place for values between 1 and 1.1.

This final dataset calculated the sample mean and standard deviation. I performed a one-sample t-test to assess whether the mean was significantly greater than 0. I also visually examined the data distribution using a histogram.

The actual mean and standard deviation of the original 1,000 draws were 0.997 and 0.982, respectively, very close to the Normal(1,1) population values as expected. However, in the final "cleaned" dataset, the sample mean was inflated to 1.181, and the standard deviation was reduced to 0.792. The t-test incorrectly rejected the null hypothesis that the true mean was less than or equal to 0 with a p-value <0.0001. The histogram revealed a distribution that was skewed right with a sharp peak just above 1 - artifacts of the data cleaning errors.

4 Impact of the Issues

Each of the simulated errors impacted the analysis in different ways:

1. The duplication of the first 100 values due to instrument overwrite reduced the adequate sample size from 1000 to 900. This decreased the precision of estimates and statistical power.

2. Converting negative values to positive substantially shifted the data distribution, inflating the mean and decreasing the standard deviation. This created a misleadingly high sample average.

3. The decimal place error affected a small range of the data but pulled those values substantially lower. However, the impact was swamped by the more significant positive bias from the sign change error.

These issues caused the analysis to reach an incorrect conclusion, as the one-sample t-test strongly rejected the null hypothesis despite the true mean being 1. The data collection and cleaning errors introduced unanticipated biases that could not be assessed from the final dataset alone.

5 Strategies for Identifying and Mitigating Errors

While it is impossible to eliminate the potential for data collection and cleaning errors, we can take several steps to minimize their occurrence and detect them when they do happen:

1. Always retain the original, raw dataset for reference. Compare summary statistics and distributions between the raw and cleaned versions to check for discrepancies.
2. Perform thorough exploratory data analysis (EDA) at every data processing stage. Visualize univariate and bivariate distributions, looking for anomalies such as shifted means, truncated ranges, or unusual peaks.
3. Implement data validation checks in the cleaning pipeline, such as range constraints and tests for duplicate records. Automate these checks and review validation output before proceeding.
4. Have multiple analysts review data cleaning scripts to catch potential logic errors. Document all cleaning steps and verify that the code matches the documentation.
5. Conduct sensitivity analyses by rerunning models and estimates on subsets of the data, such as by removing the highest and lowest 10% of values. Significant changes in results suggest potential bias from outliers or distributional issues.

Most importantly, maintain a degree of skepticism about "cleaned" datasets, especially when working with secondary data, whether data collection and processing choices could have introduced systematic biases and assessed the robustness of findings to alternative assumptions.

6 Conclusion

This simulation illustrates how easily undetected data collection and cleaning errors can introduce hard-to-identify biases that substantially impact the results of an analysis. By adopting practices such as retaining raw data, conducting thorough EDA, implementing validation checks, and performing sensitivity analyses, we can reduce the risk of these errors and improve the reliability of our findings. While no dataset is perfect, we can maximize the validity of the conclusions we draw from our data by applying scrutiny and skepticism throughout the analysis pipeline.

References

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