Navigating Data Integrity: From Simulation to Inference Amidst Unforeseen Anomalies*

Jingyi Shen

February 27, 2024

Table of contents

1 Analysis 1

Reference 2

1 Analysis

In this analytical journey, we embarked on an exploration that encapsulates the essence of data science: simulation, manipulation, and inference, all while navigating through unintended data anomalies. The initial phase involved simulating a dataset to mimic a real-world scenario where observations stem from a normal distribution with a specified mean and standard deviation. This simulation was conceived to generate 1,000 observations. However, an instrumental quirk capped the memory at 900 observations, leading to the overwriting of the final 100 entries with the first 100. This peculiar limitation mimicked potential real-world data collection issues, where hardware or software constraints might inadvertently alter the dataset.

Upon receiving the dataset, a research assistant was tasked with data cleaning and preparation, a crucial step in ensuring data quality. Unbeknownst to us, two significant errors occurred during this phase. Firstly, half of the negative values were accidentally converted to positive, introducing a systematic bias that skewed the data distribution. Secondly, an error in decimal placement altered values between 1 and 1.1, drastically reducing them. These mistakes could have profound implications on the analysis, potentially leading to erroneous conclusions. For instance, the conversion of negative values to positive would artificially inflate the mean, while the decimal place error could introduce a subtle but systematic distortion in the dataset.

[*] Code and data supporting this analysis are available at: https://github.com/CSCmaster/mini-essay-7/pulls

After addressing these issues, the final analytical step aimed to ascertain whether the mean of the true data generating process was significantly greater than 0, employing a one-sample t-test(Kim 2015). This statistical method provided a formal mechanism to test our hypothesis against the observed data, considering the accidental alterations made during the cleaning process. The t-test's outcome hinged on the assumption that the dataset, albeit manipulated, still reflected the underlying data generating process accurately enough for meaningful inference(Kim 2015).

The inadvertent errors introduced during the data cleaning phase underscore the paramount importance of rigorous data validation and verification protocols. To mitigate such risks in future analyses, several strategies can be implemented. Automated data checks can serve as an early warning system, flagging potential anomalies for review. Implementing a more robust review process, perhaps involving multiple team members, can increase the likelihood of catching errors before they impact the analysis. Additionally, maintaining detailed documentation of all data manipulation steps can aid in tracing back and correcting any mistakes, enhancing the reproducibility and credibility of the analysis.

This experience highlights the delicate balance between data integrity and the analytical objectives that guide our inquiry. The mean of the true data generating process is greater than 0.It serves as a reminder of the challenges inherent in working with real-world data, where errors and anomalies can lurk beneath the surface. By adopting comprehensive data management practices, we can navigate these challenges more effectively, ensuring that our analyses remain robust and our conclusions sound.

Reference

Kim, Tae Kyun. 2015. "T Test as a Parametric Statistic." *Korean Journal of Anesthesiology* 68 (6): 540–46.