Navigating Data Integrity: From Simulation to Inference midst Unforeseen nomalies*

Jingyi Shen

Febru ry 27, 2024

Table of contents

1 nalysis

In our analytical exploration, we embarked on a journey that delves into the core of data science, encompassing simulation, manipulation, and inference, whilst confronting unexpected data anomalies. The journey commenced with the simulation of a dataset designed to reflect a real-world scenario, where observations follow a normal distribution characterized by a predetermined mean and standard deviation. Our intent was to create 1,000 observations to closely mimic real-world data dynamics. However, a unique limitation arose due to an instrumental glitch, which restricted the dataset to 900 observations, causing the last 100 entries to be overwritten by the first 100. This peculiar constraint serves as a practical example of how real-world data collection might be impacted by hardware or software limitations, inadvertently altering the dataset and potentially skewing the analysis. This overwriting error not only reduces the dataset's diversity but also introduces a repetition bias, leading to skewed analytical outcomes and potentially misleading insights.

During the data cleaning and preparation phase, conducted by a research assistant, two significant errors were inadvertently introduced. First, half of the negative values were mistakenly converted to positive, introducing a systematic bias that altered the data's original distribution. This conversion not only skews the mean upwards but also distorts the data's true variance,

Code nd d t supporting this n lysis re v il ble t: https://github.com/CSCm ster/mini-ess y-7/pulls

impacting any subsequent analysis reliant on these metrics. Second, a decimal placement error for values between 1 and 1.1 drastically reduced their magnitude. This subtle yet significant alteration could lead to underestimation of the dataset's range and variability, affecting analyses that depend on precise value representations.

In response to these challenges, our final analytical phase aimed to determine if the mean of the true data generating process was significantly greater than 0, employing a one-sample t-test. This statistical method, intended to test our hypothesis against the observed data, assumed that despite the accidental alterations, the dataset could still reflect the underlying process accurately enough for meaningful inference. However, the introduction of the t-test without clear connection to the previous errors might seem abrupt(Kim 2015). It's important to clarify that the t-test's application here is to assess the impact of the identified errors on the central tendency of the dataset, under the assumption that the data, despite its manipulation, remains a valid representation for this specific hypothesis testing(Kim 2015).

The inadvertent errors during data cleaning underscore the critical importance of rigorous data validation and verification protocols. To mitigate such risks in future analyses, implementing automated data checks can act as a preliminary screening mechanism, identifying anomalies for further review. robust review process, possibly involving multiple team members, can significantly increase the chances of detecting errors early. dditionally, maintaining comprehensive documentation of all data manipulation steps can facilitate the tracing and correction of mistakes, thereby enhancing the analysis's reproducibility and integrity.

This exploration illustrates the delicate interplay between maintaining data integrity and achieving analytical objectives. The manipulation errors introduced serve as a stark reminder of the complexities involved in handling real-world data, where unnoticed errors and anomalies can significantly impact outcomes. dopting thorough data management practices enables us to navigate these challenges more adeptly, ensuring that our analyses are both robust and reliable, and ultimately leading to more accurate and trustworthy conclusions. Thanks Terry Tu give me valuable feedbacks.

Reference

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