Data Error Impact Analysis

Xiaoyu (Alice) Miao

2024-04-03

To provide a comprehensive analysis of the described scenario, it's crucial to delve into each step of the process, the errors introduced, their impacts on the data, and how these can be addressed or mitigated in future research endeavors. This deeper dive aims to illuminate the complexities of data integrity and analysis within a research context.

Theoretical Framework

The scenario begins with an assumption: the true data generating process is a Normal distribution with a mean of one and a standard deviation of one. Theoretically, such a distribution should reflect a symmetric, bell-shaped curve where approximately 68% of values lie within one standard deviation of the mean, 95% within two, and 99.7% within three. This theoretical distribution underpins many statistical models and assumptions, setting the stage for our expectations. Pipino, Lee, and Wang (2002)

Simulated Data Generation and Errors

In the course of generating 1,000 observations from a Normal distribution, the process encountered significant complications due to both instrumental limitations and errors introduced during data handling. The instrument used for data collection had a memory limit that inadvertently led to the overwriting of the final 100 observations with the first 100, thereby reducing the dataset's effective uniqueness and independence. This limitation imposed a repetition bias, skewing the analysis by disproportionately representing the characteristics of the overwritten observations. Furthermore, the dataset's integrity was further compromised by systematic errors introduced during the cleaning process. A notable sign error, where negative values were mistakenly converted to positive, distorted the data distribution, affecting both the mean and variance of the dataset and leading to misleading inferences about the distribution's true characteristics. Additionally, a decimal place error significantly impacted values between 1 and 1.1, inaccurately shifting their magnitude and thus disproportionately influencing a specific subset of the data. This error not only altered the distribution's overall shape

but also its central tendency measures, exacerbating the challenges in accurately interpreting the dataset's statistical properties. Taylor (1997)

Analysis and Impact

Initial Observations

The mean of the cleaned dataset was found to be significantly higher than zero, an expected result given the nature of the data generation process and the specified mean of one. However, the introduced errors complicate this interpretation. The sign error artificially inflates the dataset's mean by reducing the presence of negative values, which would otherwise lower the mean. The decimal place error, though affecting a smaller portion of the data, introduces a downward bias for those values, potentially mitigating the inflation caused by the sign error to some extent.

Statistical Testing

The use of a one-sample t-test to evaluate the hypothesis that the mean of the dataset is greater than 0 is problematic in this context. The test assumes that the data are normally distributed and that observations are independent and identically distributed—assumptions violated by the introduced errors. The repetition of the first 100 observations due to the instrument's memory limit compromises the assumption of independence, while the systematic errors affect the distribution's normality. Greenland et al. (2016)

Findings

The analysis revealed significant deviations from the expected behavior of a normally distributed dataset with a mean of 1 and a standard deviation of 1. The mean of the cleaned dataset was significantly greater than 0, which, while supportive of the original hypothesis, is misleading due to the introduced errors.

Effects of the Issues

The introduction of errors through instrumental limitations and the mishandling of data during the cleaning process profoundly compromised the integrity and interpretiveness of the dataset. Specifically, the overwriting of observations due to instrument memory constraints resulted in a dataset that did not accurately represent the intended sample of 1,000 unique observations, leading to a skewed analysis that was not reflective of the true data generating process. Moreover, the alteration of data values by the research assistant—converting negative values to positive and incorrectly adjusting the decimal place for a subset of the data—further exacerbated the issue. These actions artificially inflated the mean and introduced a systematic