Investigating the effect of Errors on Statistical Analysis*

Aamishi Avarsekar

February 27, 2024

Introduction

While dealing with real-life data, statisticians often encounter errors that could affect the accuracy of their analysis. These errors could be sourced from multiple areas like human-errors, machine-errors and unidentified bugs in code. In this paper, I would like to explore the effects of these errors while conducting statistical analysis and how we can set up measures to identify and attenuate them.Let us consider the scenario of calculating the mean of a Normal Distribution where the data is simulated with errors.

Simulating the data and required errors

The intention of this exercise is to understand the behavior of statistical analysis when errors are introduced to the data and how we can flag these unknown errors. Let us first simulate

*Code and data are available at: https://github.com/aamishi/STA302-Tutorial-07 the instrument of measurement to only be able to hold 900 unique data entries. And then we adjust the data to hold 100 more values that are a repetition of the first 100. To simulate the role of the RA, we can write code that randomly selects half of the negative draws and makes them positive. Lastly, we can also write code to divide values in the range `(1, 1.1)`. Writing code helps us maintain the true nature of randomness that could be introduced in the data. The errors are so minimal that it might not be noticeable to the human-eye.

Finding the mean of the true data generating process

As the normal distribution is determined by two parameters, mean and standard deviation, we can employ Linear Regression to estimate them as close as possible. We can also employ the method of least squares to make a comparison about our simulated mean and the actual mean. Utilizing these methods allows us to identify the potential difference between the true and errored distribution. Our goal is to calculate the mean of the true data generating process so we can declare the hypothesis of the mean being greater than 0. Using this method, we can identify the excess noise and trend-disturbance that can be introduced due to the rewriting of data and human-produced errors.

Impact of errors on analysis

The errors that can be introduced while data is collected, simulated or stored cannot always be identified before we begin analysis. When data is collected, there can human errors such as misplacing the decimal point. When data is simulated using code and technology, we can face human errors as well as code errors such as bugs that are caused by complex logic-flow. We can also encounter storage errors when the data is misplaced and stored in an incorrect format or location or limitation due to storage. While these errors might be seemingly random, they can cause the production of unnecessary noise in the data or can completely skew the results of the hypothesis. In the errors that we introduced in our data generating process, the repetitions of the first 100 values can cause a bias in our analysis, gave us incorrect readings for trends and also cause errors in estimating the mean and standard deviation of the distribution.

The RA's human errors that were accounted into the simulated data could also bring additional noise and influence the unbiasness of the data. If we are dealing with data that are close to 0 and we want to test our hypothesis of the mean of the true data generation process to be 0, this could vastly affect the reliability of our analysis and produce highly skewed results.

Identifying errors

Before conducting analysis

Before we begin work with real data, we should estimate the nature of the data and the general shape of the distribution. We can simulate our own data based on facts of the hypothesis. For example, if we are measuring the height of high school students, we can look into the age, sex and racial demographic of the students, and expect a height distribution that is in proportion to these factors. Any anomaly, such as the average height being close to 130CM for students aged 16-18, can be in an indication that there might be errors in the real data and our simulated data can serve as evidence for this.

We can also simulate data for a large number of repetitions. This would bring out a distribution that is close to the natural distribution of the data following the law of large numbers. If our distribution after conducting the analysis does not have this general shape, it can also be indicative that there could be an error in the data.

While conducting analysis

Certain errors like bugs in code should be thoroughly tested. We should set up several code invariants that ensure the general expected behavior of the data and the code. We can also write tests that must pass in order to proceed to the next part of the code.

We can also try to simualate the data by taking a subset of the real data, and produce a data set of the same size by allowing repetion of the data. We can flag inconsistent trends if the there are unusual spikes in the values of the mean or median.

Conclusion

Through the exercise of this paper, we learn that before we begin any analysis, we must always account for the possibility of errors in the data set and not just the actual procedure. Setting up thorough methods of identifying the errors will make our analysis more reliable and produce more accurate results that are true to the nature of the expected distribution.

However, we must also take into consideration that not all errors in the process of generating and storing data can be identified and have appropriate mitigation strategies. To solidify our findings, we would have to look into other strategies of hypothesis testing and compare if we have overall uniform results.

Peer Review

I would like to thank Chay Park for her review to make improvements to my paper. You can find her review at this link: https://github.com/aamishi/STA302-Tutorial-07/issues/1. Her student number is 1006486982 and email is chay.park@mail.utoronto.ca.