A simulation and analytical study of a normal distribution with many unexpected errors.[*]

Yunzhao Li Wentao Sun

February 27, 2024

Simulation

```
[1] 1.165457
```

```
[1] 0.8442421
```

What I did

I simulated the following situation in order with R code (R Core Team 2023):

Allow that the true data generating process is a Normal distribution with mean of one, and standard deviation of 1. We obtain a sample of 1,000 observations using some instrument.

1.Unknown to us, the instrument has a mistake in it, which means that it has a maximum memory of 900 observations, and begins over-writing at that point, so the final 100 observations are actually a repeat of the first 100.

2.We employ a research assistant to clean and prepare the dataset. During the process of doing this, unknown to us, they accidentally change half of the negative draws to be positive.

3.They additionally, accidentally, change the decimal place on any value between 1 and 1.1, so that, for instance 1 becomes 0.1, and 1.1 would become 0.11.

4.You finally get the cleaned dataset and are interested in understanding whether the mean of the true data generating process is greater than 0.(Alexander 2024)

[*] Code and data are available at: https://github.com/yunzhaol/Simulation_accidents.git

What I found

I found the mean is 1.165457 which is greater than 0 and 1. And the standard deviation is 0.8442421 which is less than the deviation of 1 in the assumption of true data generating process.

What effect the issues had

They made the mean higher and the standard deviation lower. They modified some values of the draws. They made final 100 observations a repeat of the first 100, half of the negative value positive and change the decimal place on any value between 1 and 1.1. The effect above made changes to the data we draw in different ways.

What steps I can put in place to ensure actual analysis has a chance to flag some of these issues

Compare the mean and standard deviation I get to the one in assumption. If they are different, somewhere must go wrong. Check the value of the draws to see how it changes as each step move on. For example, if some negative values in draws turn positive after cleaning process, we will know mistakes happened here. We can achieve this with union function. By creating a union of the draws before the particular step and the one after the step, then find the difference from showing the draws in the union but not in the one before or after the step. Doing this to each step, we can flag the issues.

Reference

Alexander, Rohan. 2024. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com.

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