Problems in practice: errors in data collection*

Jacob Gilbert

Introduction

When doing data analysis it is not uncommon for mistakes to be made. This can be accentuated when data is going through many hands, as well as problems with the hardware or source of the data. The following is a simulated example of how problems in the data generating process can lead to bad analysis. 1,000 data points were generated and to simulate some common problems in data collection changes were made. The data was created in R (@citeR) using packages tidyverse (@citetidy) and knitr (@knitr).

Originally this data was created by a random sample of points from a normal distribution with mean and sd of 1. As the sample is somewhat substantial in size the estimated mean and standard deviation from the sample are very close to that of the population distribution.

Estimated mean	Estimated standard deviation	Estimated standard error	P-value
0.97	0.98	0.03	0.32

Figure 1: Summary statistics of the original distribution

Simulation

The changes made to the sample are to simulate the following real life situations. - A problem with the data collecting instrument that only has memory for 900 observations. As a result of this the first 100 observations are repeated. - Human error led to half of the data points below 0 being converted to their absolute value. - Any number between 1 and 1.1 was divided by 10.

*Github link: https://github.com/JfpGilbert0/Simulated-data-collection. thanks to Alexander Guarasci for review

Expectation

We would expect these errors to have mixed effects on the analysis dont on the dataset. Firstly the repetition of data points could have unpredictable effects on the mean and standard deviation. The impact will be determined by what is in the repeated data. If there are extreme points in the first 100 elements then the changes could be great, if the points are mostly around the mean then this cause the opposite bias. The effect of changing the sign of negative is clear. This would cause bias towards the correct mean and decrease the observed standard deviation. The final change causes numbers originally less than 0.1 away from the mean, the other way, by 0.9. In this simulation there are 46 such occurrences, almost 5% of the sample, thus this change is expected to be significant.

Results

Estimated mean	Estimated standard deviation	Estimated standard error	P-Value
1.02	0.9	0.03	0.34

We see that both the mean and the standard deviation have changed. Due to the conflicting effects of the different collection errors it is unclear which is the strongest driving force for this change. Looking at the P-value we see that the confidence that the population mean is 1 has fallen from the original. This is a case when the biases are moving us further away from our hypothesis than is true. This means conclusions being taken are incorrect and is problematic.

solutions in practice

There are many ways to avoid problems such as these. Using the correct tools for the job is important. Underpowered hardware is a problem that is easily avoided by bringing something stronger than you need, this is not always possible however should be done when it is. Creating an excellent chain of custody with data is a great way to catch problems and find where they are originating. DBT is a great solution to this as it creates a solid data cleaning structure from raw to cleaned. The reason this is useful is when problems occur you can get quick flags being raised and 'co-ordinates' for the problem area. This is reliant on having appropriate tests at every point of the data cleaning process. Testing is at the heart of having clean usable data, this can create a strong safety net for common problems.