Parsh Gandhi

27th June 2022

ATDP secondary division

Data Science Lab with R & Python, Section 1

Reflection #2

From the second chapter of *Practical Statistics for Data Scientists* the main things I learned were samples and how the size and quantity of them affect the data and outcome but also bias that gets built in the system along with standard error. Samples allow a smaller dataset to be taken from a bigger population. This can offer the choice to choose quantity or quality of the data to be taken. Choosing quality often gives more accuracy and a lot of different data to be added. Choosing quantity often gives less diverse data but has an opportunity to explore more dense data. I like this book as it helps me understand the concepts not just with the thinking behind it but with real world examples. One example the book gave for using smaller samples was in the 1936 election polls to decide who would win. Using these smaller samples the polls encapsulated many different individuals and correctly predicted the win of Franklin Roosevelt but on the other side the polls that used a bigger population were less diverse and therefore a major chunk of the population voted one person rather than being spread out leading to an incorrect prediction. One example for the population data was google's search engine where we can find anything we want at the type of a few letters. But what I have a question about regarding google's search engine is, when searching through big amounts of data why is it sometimes less accurate to put in a very long and precise search rather than a shorter one? Why does this occur if the sample size for the words get smaller, shouldn't it be more accurate then? Another thing I learned in this section was the bias that can occur through the samples that we supply to get data from and from inside the system itself. The example for this was the same poll on where they took the whole population and this would skew the bias as it is only accounting for the people who own luxuries and are their own subscribers. Standard error is also linked to this but instead as we have a larger sample or population the error goes down. These are the main points I took away from the second chapter reading of *Practical Statistics for Data Scientists*.