**Reflection before Class Three**

As I read through the first half of the second chapter of *Practical Statistics for Data Science*, I learned a lot about why bias is often seen in data and ways that this can be prevented. This is an extremely important topic. As I stated before, bias in machine learning algorithms due to prejudice in the data fed into those algorithms can cause a lot of trouble. Many people face scrutiny for factors irrelevant to the valid reasons for which a person should face scrutiny. In the *Pre-Course Reflection*, I stated, “This is a topic that has come up fairly often in the past few years, as recently, there have been a large number of cases where minorities have faced prejudice due to factors they possess that they have no control over but are frowned upon. This has caused a lot of people to face scrutiny when the only thing that made them stand out was, for example, a physical trait they had, such as their skin color.”

There were a number of concepts that were discussed within this half of the chapter. I was not familiar with all of them, but they all seemed to make sense. The first term I encountered was random selection (or random sampling), which is the method of sampling that helps data scientists avoid sample bias (note the word ‘random’ within the term, as that word tells us how this method of sampling works). However, this isn’t as easy as it may seem. In order to sample unbiased using random selection, data scientists must first figure out what kinds of data fit within the population to be sampled. In essence, the population must be free of bias for the sample to be free of bias.

Using the example of bias mentioned in the first paragraph (bias due to skin color), it is likely that this bias would be minimal if a random selection was done from each of the ethnic groups that made up the dataset with an equal sample size from each group. This way, the sample bias that existed would be due to problematic data rather than because most of the data is from a majority group, which would result in the population (and thus, sample) mean becoming skewed towards data from the majority group. This idea of dividing the population into strata (homogeneous subgroups of a population with common characteristics) and randomly sampling from each one is called stratified sampling.

I was already aware of selection bias and the often-seen regression to the mean when taking successive measurements of an outlier, but I was surprised to learn about the central limit theorem. I did not realize that means drawn from multiple samples would more closely resemble the bell-shaped curve than the source population. This seems like an important fact to know, and I will be keeping this in mind in the future.

In terms of the first class, I was pleasantly surprised. I had originally been afraid that the class would be either too simple or too difficult for me to keep up with, but it turned out that I was familiar with some of the content and unfamiliar with some of it, which meant that I kept learning throughout the class. Although I can’t remember which definitions I learned off the top of my head, I can say that I did understand them as well as the concepts we covered in class. I think that the pace for this course is just right; it gave me time to finish what I needed to as well as a little extra to double-check what I did.