DSC 530

Kiana Gonzalez-Rodholm

6/3/2021

For my project I chose to work with a data set containing information on COVID vaccine distribution for various countries. My statistical question was: how does the United States compare to other countries as far as vaccine distribution? I also wanted to find out how the variables were related and find general information from the data by country. I found through my analysis that a lognormal model is a good fit for the distribution of this data. I also found that although slow to start, the US has exponentially grown in vaccine distribution compared to other countries. When comparing different variables together to find correlation, I discovered that total vaccinations and people fully vaccinated are positively correlated and statistically significant. This is to be expected because the higher number of people that are fully vaccinated means that more people had one or two shots, and therefore contribute to the total number of vaccinations.

I felt like there were a few things missed in my analysis. If I had more time to clean the data, I could have spent more time removing outliers. While I removed Nan values, there were a prominent amount of zero entries which I think skewed the data a bit. I think if the data were updated more recently, it could potentially show different results, especially with the US data. Since this data was collected at the start of the term, it is slightly outdated with the rise in vaccine distribution over the last few months. I do not think there were any variables that could have helped the analysis, however if there was more time to plug through all the time series data then I could have potentially used date and time as a useful variable as well.

I do not think I made any incorrect assumptions based on the results I was seeing, however as mentioned above, it is possible that the data could be skewed by the number of zero entries. Some additional challenges were making sure the textbook functions were being used properly, as the course was so heavily reliant on the textbook package thinkplot2. I wish we had the option to use more standard libraries rather than just the code and functions provided by the text. I also feel like I still do not fully understand how PMFs and CDFs fully help us understand our data. I wish there were more explanation in the text or simpler examples to make understanding these functions a bit easier.

I did enjoy the visualization aspect of this project, as well as applying all the things we have learned in the course. I also enjoyed displaying the information on a PowerPoint which helps the audience see the results better. I look forward to more data analysis projects like this one in the future.