**Summer Olympics**

Is there a home field advantage?

My goal was to quantify the likelihood of a country winning more medals at the Summer Olympics while playing the host country. I don’t feel comfortable saying that I have an exact amount, but the statistical calculations – particularly the Probability Mass Function - indicate a slightly higher likelihood of winning more medals while hosting if the number of medals won overall is also high.

What do you feel was missed during the analysis?

Due to time constraints, I was unable to perform further analysis on how the total number of medals won affected the likelihood of winning as a host. While there were some indications towards this conclusion, there is opportunity for further analysis.

Additionally, I would have liked to have the ability to explore the role that gender plays in the statistical modeling and calculations. However, for simplicity in research and desire to meet deadlines, I did not explore this further.

Were there any variables you felt could have helped in the analysis?

The data set that I found was for all medals won by all countries from 1976 to 2008. While this still encompassed a large amount of data, it would have been helpful to have the information on all individual athletes that competed, including the ones that did not win. The countries that had high numbers of medals one also tended to have larger populations, and therefore larger Olympic teams. It could have shed more light on the games as a whole, rather than just the instances of winning.

Were there any assumptions made you felt were incorrect?

I attempted to approach this with little or no assumptions. With that being said, I did not realize the sheer number of individual events included in the Summer Olympics. I was aware of the “popular” sports, such as Gymnastics and Swimming, but didn’t realize that hockey was included in the Summer Olympics. Additionally, I was not aware of how many individual swimming events there were. There are so many that they make up 14% of all events.

What challenges did you face, what did you not fully understand?

The original dataset did not include a few variables that I was able to create. However, even with my new variables, all of the data was categorical. There were no variables with numeric data (discounting the Year, as it was categorized as categorical). This created a challenge to create many new dataframes from the existing data with new variables that were numeric, in order to perform simple calculations like mean or create scatterplots.