Statistical Analysis and walkthrough of Rugby Union Data

Introduction:

In this project we will investigate a data set for Rugby Union matches. In this data set, we have 10 International teams that fall under the Rugby Union umbrella and have been competing against each other since 1871. Each of these teams have played against each other several times either in their respective home venues or in neutral venues during world cup competitions. The data set contains several variables like the stadium they played in and the type of competition which we will ultimately omit. We will look at certain matchups between teams in a Shiny App which reads in a csv file that contains all the Rugby Union results.

Goal:

Ultimately, the goal of this project is for a person interested in analyzing historical results between Rugby Unions teams to have an informed understanding of what historical results can tell us about the data. We will investigate score distributions in order to make a fair estimate of which team you think might be more likely to win based on historical results.

As part of the project, I will use an example to work through as part of the analysis in order to guide through questions of the analysis.

Part 1:

After running the shiny app, in the drop-down menu to your left, select two teams you would like to investigate. In this example case, I will choose South Africa as the home team and Wales as the away team and hit analyze match.

Before scrolling ahead, firstly assess and understand the two density graphs to your right. The first Density should show South Africa’s score density of all their matches. This density graph shows a distribution of all of South Africa’s scores they have achieved throughout history against all the Rugby Union teams. The peak of the density is ~ 20 points which in essence means that this is the most common score South Africa will achieve. The density graph can easily be understood just that the higher points in the graph are more likely than lower points. That is clear because extreme and high values like 80 + are very small and negligible in likelihood.

The same principles apply to the Wales density but will of course look different due to differing historical results in all their Rugby Union matches.

Now, consider the third density graph, which is a density graph of historical scores when these two met. It only considers the matches that these two played against each other. Notice Wales’s and South Africa’s densities. Wales has a big peak around 15 points and sharply decreases after that. South Africa has a peak spanning more horizontally and almost mimics a uniform distribution, meaning that the probability that South Africa score between 20 and 40 points against Wales is almost the same.

Q1: Are you able to, based on the three density graphs of South Africa v Wales, predict who is most likely to win between the two?

Answer: Although these distribution graphs are useful (especially the third one), we can’t truly make a factually based assumption about who would win, but we can guess by looking at this third density graph. Based on the observation that South Africa is more likely to score ~ 22 + points against Wales and has a tail which is more gradual, we can assume that South Africa has historically scored higher points against Wales than Wales did. So, we can assume that South Africa will most likely win match up against each other.

Part 2:

Now consider the point differential distribution of South Africa vs Wales. This graph can be understood where values far to left and to the right (away from 0) are games between South Africa and Wales where one team scored significantly more points than the other. Values to the left of 0 represent games Wales haven won and to the right of 0 represent games South Africa have won.

Q: After considering the point differential distribution and the densities, can you predict who might win?

Answer: Yes. Combining all the visuals so far, we can again say with confidence that South Africa is the more likely winner in this matchup. The point differential distribution clearly leans to the right of 0, meaning there have been more games where South Africa has outscored Wales. While there are a few matches on the left side of 0 (Wales wins), they’re not only fewer in number but also tend to be by smaller margins. This adds further support to what we saw in the density graphs — that South Africa has generally had the edge when these two teams have played.

Part 3:

Now notice the Pie chart agrees with us! The Total Score vs. Point Differential we have an interactive graph where we can see the final score of each individual match. Points the right represent games with large match score whether that be either or both teams. Points along the Y axis (up or down/above or below 0) represent the point difference.

Q: Based on the graph, can you point out 2 different points on the graph that look interesting to you and report their score.

Answer:

* The point to the far top right, SA 96 – Wales 13. This point is an outlier of a game with a huge point difference and a huge match score. This applies to our thinking that points far above or below the 0 on the vertical represent games that are largely one sided.
* The point to far right very close to the 0 horizontally, Wales 36 – SA 38. This point represents another important aspect of our graph. Not only does it show that it was a large scoring game by being far right on the graph, but by being almost on the 0 horizontally, it is also a very close game. In theory this should be a very entertaining match because it was such a large scoring game and was very close!

Part 4:

Now finally, just to conclude, we can see some summary statistics and other information that can confirm what we estimated earlier on as well as a table with all matches between our two selected teams.

Feel free to now select other team, investigate other match-ups and study other densities and continue to challenge your statistical knowledge by analyzing visualizations and attempting to get to conclusions based on them!