Data 2010: Progress Report

Owen Ostermann, Daniel Paul, Tom Rodov

March 20, 2023

Section 1: Data Analysis

There are three central aspects of the data that we are interested in understanding. The total amount of goals scored in a game, the total amount of goals scored by a certain team, and the difference in goals over all games in the World Cup. We will provide some tables that include some summary statistics about these three categories.

The total amount of goals scored in a game between both teams playing (over all games in the dataset):

mean	median	sd	min	max	IQR
2.83	3.00	1.93	0.00	12.00	3.00

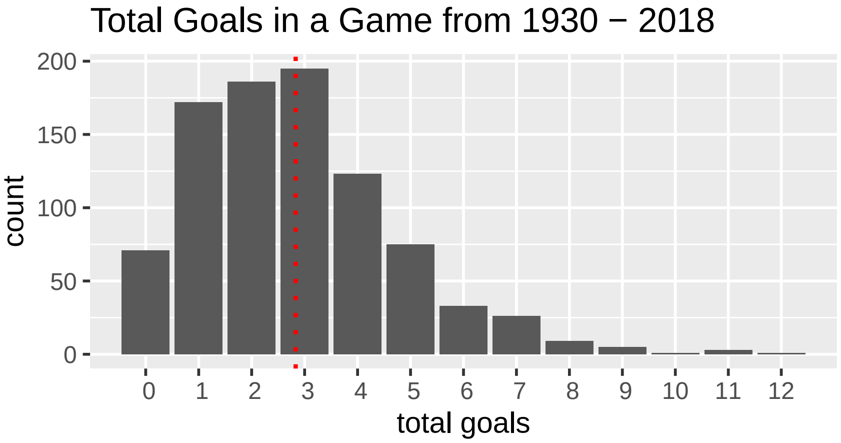
The total amount of goals scored in a game by France (over all of games France participates in):

mean	median	sd	min	max	IQR
1.82	1.00	1.61	0.00	7.00	2.00

The difference in goals between the two teams playing (over all games in the dataset):

mean	median	sd	min	max	IQR
1.50	1.00	1.42	0.00	9.00	1.00

We show an example of a bar plot of the sample data explained above:



Now we should look at some interesting correlation values between different parts of the data. First we could look at the correlation between the total goals scored and the year of the World Cup Tournament. When computing this value we get **0.8487671**. This number gives us some interesting information that there is a trend upwards over the years, meaning in more recent World Cups, more goals are being scored than in older World Cups. Next we calculate the correlation between the number of goals scored by a team in the match and the result of the match. This is to see if there is a relationship between winning and scoring more goals in a match. After filtering the dataset to show only the amount of goals scored (we add a row for both the home and away team in the game) and the result of that corresponding game (win, loss, draw), we get a correlation value of **0.6404849**, implying that the more goals scored is associated with a team winning the game. Finally what if we observe the total goals scored and the attendance of fans at the World Cup, do the fans have an influence on how the teams preform? Do more fans mean more exciting games overall? When computing this value we get **0.7877445**, implying that the more fans in attendance, more exciting the matches will be as more goals are scored.

Section 2: Current Status of the Project

As of the current status of the project, we have made significant progress in analyzing the FIFA world cup dataset. We have collected data regarding the distribution of the total number of goals scored by teams, the difference in goals in games, and France's goals over the years. After analyzing the data, we have determined that these variables follow a Poisson distribution. The Poisson distribution is a discrete probability distribution that describes the given number of events occurring in a fixed amount of time or space. In the context of soccer the Poisson distribution is a great model for goals scored in a game as they are rare events that occur within a timed interval, the length of the game. The Poisson distribution is characterized by a single parameter λ, which represents the average amount of events occurring in that timed interval. With the historical data we are given in the World Cup dataset, we are able to predict the λ values of teams and then use the Poisson distribution to generate the likelihood of teams scoring a certain amount of goals in a soccer game.

To confirm our hypothesis that goals in a soccer game follow a Poisson distribution, we have conducted a Kolmogorov–Smirnov test with a 95% confidence level. For example, we analyzed the total amount of goals scored in a game by mutating the dataset and adding a column called total goals in a game. We then filtered the data to generate a data frame with just the amount of total goals scored in a game. We calculated the mean value, or λ value of the sample data. Taking that λ value we generated the same amount of random Poisson variables from a distribution with the same λ value. We Plotted the data from World Cups and the randomly generated sample and noticed that they followed a substantially similar distribution. Our null hypothesis was that the two samples are from the same distribution, the alternative hypothesis being that there is not enough significant evidence to say they are from the same distribution. While running a KS test at the 95% confidence level of these two samples we found that the p-value was much larger than 0.05, thus we fail to reject the null hypothesis and we can say that the two samples come from the same distribution. Since we tested our data from World Cups against a randomly generated Poisson distribution, we statistically concluded that World Cups data of total goals scored in a game is from a Poisson distribution. We repeated this process for France's goals scored over all games and the absolute value of the difference in goals scored between the home and away team, getting the same results.

In conclusion, we have made substantial progress in the FIFA world cup dataset analysis, and we have confirmed that the data follows a Poisson distribution. This confirmation will help us to identify patterns and make accurate predictions about team performance.

Section 3: Future Direction of the Project

The fundamental objective of our project is to create a predictive model that can provide sports bettors with more information about goal spreads in the FIFA World Cup. Specifically, we aim to develop a comprehensive probability table that will feature home versus away goals as the primary variables of interest. The table will assign a probability to each conceivable pairing of goals scored and conceded, generated using a hybrid approach that integrates regression and Elo ranking weighting. This combined framework is designed to improve upon the limitations of traditional regression and non-weighted probabilities, creating a more robust and reliable model that can be evaluated against a baseline regression model.

Our algorithm will take into account numerical data about goal spreads, which refer to the difference between goals scored and goals allowed, a statistic that sports bettors often wager upon. The algorithm will have access to this information from each game in the FIFA World Cup history allowing countries with a successful history to be recognized by the model. By utilizing this data, we aim to streamline the process of making informed betting decisions. Our model's efficacy will be evaluated through comparison tests to determine if it produces quality output that can be utilized by pragmatic bettors. Overall, our goal is to create a predictive model that can provide sports bettors with a greater degree of information and help them make more informed betting decisions.