
ISyE 2028 – Basic Statistical Methods - Fall 2015

Bonus Project: "Big" Data Analytics Goals Scored in Europe's Top Leagues Ryan Sanders

For this project, I would like to analyze the number of goals scored in the Bundesliga(Germany), Serie A (Italy), Premier League(UK), and La Liga(Spain). In the soccer world, there is a lot of debate over which league is the "best". The goal of this assignment is to determine which league has scored the most goals over the past few seasons and during which time of year the teams in the respective leagues score most frequently. I am interested in how the time of year and busyness of schedule affects the domestic leagues. Throughout the season, the teams compete in their respective leagues but some of the players are called to play in international competitions for their national teams. So with this in mind I will look at when the international breaks are, as well as when other competitions are going on and how this might affect the leagues performance in terms of goals scored.

To get my data I will go on the sites of the Leagues and look at the results for every team and calculate the total number of goals scored each month. I will then take those numbers and look up the dates for the international breaks and other competitions keep the data from goals scored in each of the months and then the dates of the other competitions in a spreadsheet or chart. I will see which league scored the most goals total, and also which month each league scored the most. Based on this data I will then be able to see how the breaks either increase or decrease the goals scored. It will also tell me which league is "best" by seeing how each league scores goals and the frequency of these goals.

Once I collect this data, I will be able to construct confidence intervals to determine the range of goals to expect for each league given the past few seasons. Also, given that each season is almost halfway over for this year, I will be able to predict or determine how many goals will be scored in the second half of each season. With the given size of the data (months in which each season spans for four leagues), I should be able to utilize the standard normal Z table taking into account that I do not know the true population variance. I predict that my data will show me that the winter months are when each league scores the most, and that the beginnings of each season are when the least amount of goals are scored. With all this data in mind, I will be able to determine which league is "best" in terms of goals scored.