PHSX 815 Project 2: Analyzing a Soccer Team's Performance: A Slightly More Realistic Simulation and Hypothesis Testing of the Number of Goals Scored by a Team Per Game

Ashley Lieber

March 11, 2023

1 Introduction

As the year 2022 drew to a close, much of the world had their eyes on the television screen to see which country's soccer team would rise to the top and win the FIFA World Cup. This event captivated millions across the globe and serves as the inspiration for the simple simulation described in this paper. As a contrast to the work presented in Project 1, this work aims to draw closer to reality by introducing more complicated components in generating our simulated data. This project gives credence to the fact that a team's performance in a single game can be dependent on a multitude of factors. In this case, instead of simply sampling the data straight from a given distribution, there will be a set of prior parameters used to generate the distributions the data is sampled from. While it only takes one step further in complication, it is trending in the right direction towards reality.

This paper follows the situation in which one would observe many soccer games played by a single team over many seasons and record the number of goals scored by the team for each game. The main difference between this project and the previous is in how the data is generated, namely the probability distribution it is sampled from. In the previous project, the simulated data was sampled from a Poisson distribution that was generated off of one given rate. In contrast, this project aims to complicate that scenario. A gamma distribution, the conjugate prior of the Poisson distribution was used to generate the Poisson distribution the goal data was sampled from.

Nevertheless, the analysis of this data will be very similar to that done in project 1. In this set up, two different data sets, corresponding to two different hypotheses, can still be generated and hypothesis tested. The ultimate goal in this analysis still remains to answer the questions, how well can the two different hypotheses be distinguished from each other? If we were dealing with real world data, we would be able to distinguish which rate or set of prior parameters is more reflective of the actual data set or the team's performance.

This paper is organized as follows: Sec. 2 explains the different hypothesized scenarios that were tested. Sec. 3 describes the Python experiment simulation that was developed to simulate the data for these hypotheses. Next, Sec. 4 explains how these outputs were analyzed and interpreted. Next, Sec. 5. Lastly, the conclusions of this simulation are presented in Sec.6.

2 Hypotheses Explanation

For the purposes of this paper, two different sets of hypotheses were chosen to be tested: 1. Separated Hypotheses, and 2. Overlapped Hypotheses. The first set of hypotheses, aptly named Separated Hypotheses, were chosen because throughout testing the analysis code, it appeared that even hypotheses with relatively similar parameters could be determined to be too dissimilar to do meaningful hypothesis testing on. Based on a given data point, the analysis could definitively tell you which hypothesis that observation did not come from. The second set of hypotheses, named Overlapped Hypotheses, represent the opposite situation to the previous set. These hypotheses are even more similar so that meaningful hypothesis testing can be conducted on them to differentiate between the two.

With the big picture of the two sets of hypotheses in mind, we will now discuss what parameters were chosen and why. For each hypothesis, a data set needs to be generated which consists of many experiments that each have many single game observations within them. In the soccer scenario, this is most akin to observing many seasons of a teams performance in which each season has many games. As seen in the previous project, the more data is gathered, the more accurate and precise the analysis will be – of course to a limit of statistical significance. Instead of spending the time and effort to record this data, the code in this paper can simulate the distribution of data that we could expect given certain parameters. In the previous iteration of this simulation, the user was able to set the rate that they wanted to test for the team and then data was generated based on a Poisson distribution. In this simulation, the data generation is slightly more complication. Rather than the user inputting a guess at the rate that the team scores per game on average, the user inputs values for alpha (α) and beta (β). These are the configurable parameters for the Gamma distribution which correspond to the size and scale (inverse beta) of the distribution. From this distribution, a value can be randomly sampled which will serve as the rate, denoted by the Greek letter λ, from which the Poisson distribution will be generated. Then the measurement will be sampled from a Poisson distribution for that rate. This sampling of the rate is done for each measurement. These data sets that were generated for each hypothesis can be viewed in the GitHub repository linked in Section 7. Additionally, the gamma distributions for each hypothesis pair can be manually compared using the $GammaGraphComparison.py$ python file.

The parameters that the Separated Hypotheses were generated based off of are detailed below:

1. Null Hypothesis H0: $\alpha = 2$, $\beta = 1.5$
2. Alternative Hypothesis H1: $\alpha = 4$, $\beta = 2.5$

These parameters were chosen because the distributions they represented, shown in Fig. 1, have markedly different means and variances (spreads). This provided an interesting scenario to hypothesis test and study whether they could be distinguished.

The parameters that were used to generate the Overlapped Hypotheses are detailed below:

1. Null Hypothesis H0: $\alpha = 2$, $\beta = 1.5$
2. Alternative Hypothesis H1: $\alpha = 3$, $\beta = 1.1$

The gamma distributions that these hypotheses represent are shown in Fig. 2. These distributions are obviously much more similar than the previous set of hypotheses. These two hypotheses describe gamma distributions that have different means ($\alpha = 2, \alpha = 3$), but very similar spreads.

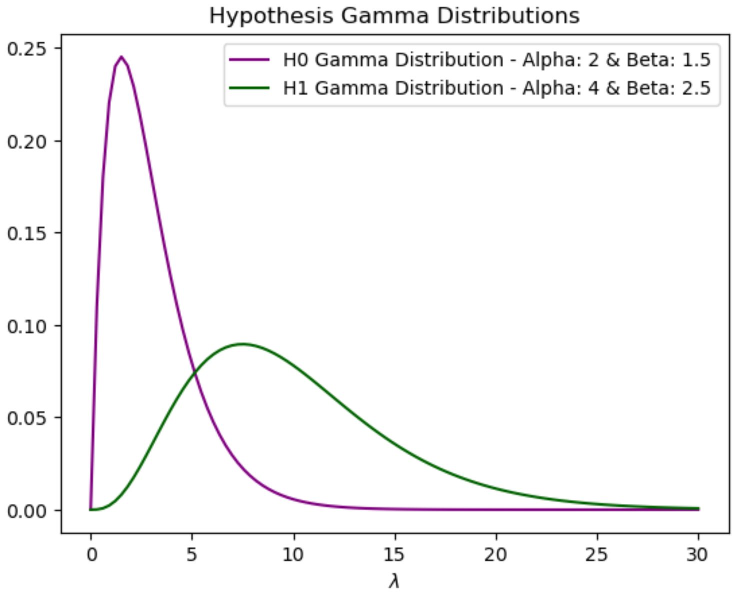

Figure 1: A plot of the gamma distributions for the Separated Hypotheses. These hypotheses were initially tested because their means and variance were markedly different.

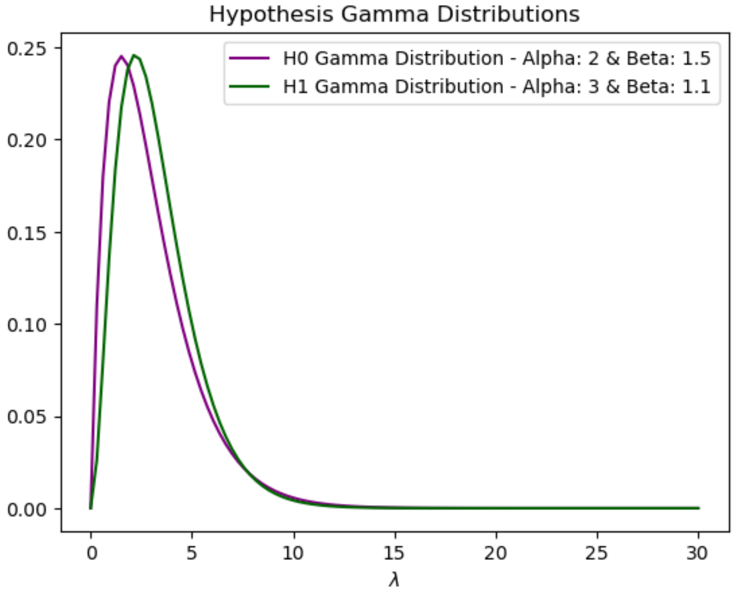

Figure 2: A plot of the gamma distributions for the Overlapped Hypotheses. These were chosen because the distributions are much more similar

With these hypotheses in hand, the code will generate a data set for each hypothesis and subsequently analyze this data which will aid in determining which hypothesis is more likely to reflect the true average for the soccer team in question.

3 Code Generation and Experimental Simulation

In order to conduct this analysis, a data set for each hypothesis is needed. The goal was to use a prior distribution with a set of configurable, nuisance parameters that can be used to generate a Poisson distribution to sample a data point from. Due to this goal, the data in this simulation is dependent on a Poisson-Gamma model. A gamma distribution was the obvious and convenient choice since it is the conjugate prior of the Poisson distribution which "means that we can actually solve the posterior distribution in closed form," [1]. The gamma distribution is dependent on three parameters, the threshold parameter, the shape parameter (α), and the scale parameter (β). The equation below shows the probability density function for the gamma distribution for when $x > 0$, and $\alpha, \beta > 0$.

$$f(x|\alpha, \beta) = \frac{x^{\alpha-1}e^{-\beta x}\beta^{\alpha}}{\Gamma(\alpha)} \tag{1}$$

where $\Gamma(\alpha)$ is the gamma function. For all positive integers, $\Gamma(\alpha) = (\alpha - 1)!$ [2]. The domain of the gamma distribution is generally $(0, \infty)$ where the parameters $\alpha, \beta > 0$.

The threshold parameter sets the lowest limit of the distribution which is generally only utilized in order to allow the distribution to handle negative values [2]. For our purposes, the threshold was set to zero for all hypotheses since we only wanted to consider positive values. The shape parameter (α) generally "specifies the number of events you are modelling," and the scale parameter (β) is the mean [2]. Oftentimes, an alternate version of the scale parameter is used called the rate (λ). These two values are simply reciprocals of one another as shown below: [2].

$$\beta = \frac{1}{\lambda} \tag{2}$$

$$\lambda = \frac{1}{\beta} \tag{3}$$

In this case, these parameters were chosen, as described in Section 2, by visually inspecting the resulting distributions and choosing spreads as I desired [3]. Further experiments may put more thought behind those decisions.

When generating the data set the user will be able to enter their desired values for these configurable parameters. Those values set that gamma distribution that will be used as the prior distribution for the Poisson sampling. Additionally, the user can also set other values such as those described in the list below.

1. Nmeas: number of measurements to be taken per experiment. This refers to the number of games we wish to observe per season. One data point is the number of goals scored in a single game.
2. Nexp: number of the experiments. In this context, this can be thought of as the number of seasons with Nmeas number of games. For example, if the user sets $N_{meas} = 20$ and $N_{exp} = 10,000$, then the code will generate 200,000 data points which simulates the observations of 200,000 games.
3. Seed: the desired seed value
4. Output: The filename for the data to be saved to. If not specified, the data will output to the command line.
5. Rate Output: The filename used to save the list of all of the rates used which were sampled from the gamma distribution.

For each measurement that is taken, a value will be sampled fromt he given gamma distribution which will be a real number but not necessarily an integer. This value will then be used as the rate (λ) to set the Poisson distribution. From this distribution, a value will be sampled which will be the data point used to simulate the number of goals measured for that game. Even though the lambda value sampled from the gamma distribution need only be a real number, the observable (number of goals scored) will be an integer since it is sampled from the Poisson distribution. This process is done over and over again to generate the data set. The same gamma distribution is used for all of the measurements, but every rate and subsequent data point is randomly sampled.

For completeness, the rest of this section will explain why the Poisson distribution, from which the data is sampled, was a viable choice for this simulation. This distribution was carefully chosen because the characteristics of our scenario meet the criteria for a Poisson distribution. These criteria are as follows: (1) the "individual events occur at random and independently in a given interval (this interval can be of time or space," and (2) "the mean number of occurrences of events in an interval (time or space) is finite and known" represented by the symbol λ [4]. This particular experiment fits these criteria because the measurements of goals scored by a team are random events recorded once per game and each game is in and of itself an independent event. Additionally, the average number of occurrences, λ is a finite number and known under each hypothesis. Furthermore, the Poisson distribution is a "discrete probability distribution that describes probabilities for counts of events [5]. Thus, the Poisson distribution is an apt choice to randomly sample our data sets from The Poisson Probability equation is given by the following formula which calculates the probability of x occurring in a given interval

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{4}$$

where λ is the average number of measurements per interval, e is the constant Euler's number which is approximately 2.78, and x which takes discrete values and corresponds to our data [6].

At the end of the data generation step, a plot will be made that shows the gamma distribution according to the user-inputted parameters as well as a histogram of all of the rates that were sampled throughout the course of the experiment, which should in theory represent the same curve. For example, in Figure 3 this comparison can be plainly seen. Since a large number of measurements were simulated, the sampled rates appear in the same distribution as the gamma from which they are sampled.

The data that is simulated for each hypothesis is then sent to another Python script which is titled $GoalDataAnalysis.py$ which further analyzes the data and performs the hypothesis tests needed.

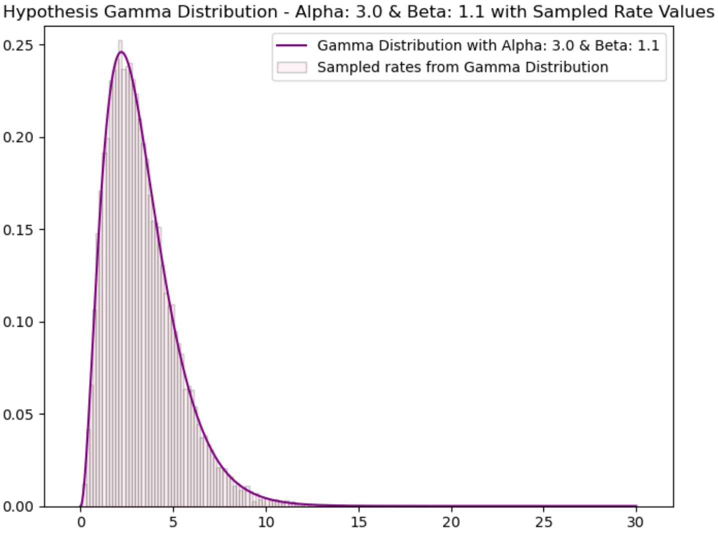

Figure 3: A comparison of the gamma distribution that results from the user inputted α and β parameters along with a histogram showing all of the values that were sampled from the gamma distribution over the course of the experiment.

4 Analysis

Now that the data sets have been generated, we have completed the simulated of many single game outcomes under each hypothesis. For one hypothesis, the vector of outcomes will read. For the first hypothesis $H0$, this vector will be called $X0$, and similarly, for the alternative hypothesis $H1$ the data vector will be called $X1$. These vectors of data are then put into a histogram with many bins – using integer binning. The binning is done such that the i-th bin (i.e. how many times "i" goals were scored) is filled with the quantity "Ni." This is set in up in away such that the following equation holds.

$$\sum_i N_i = N_{total} \tag{5}$$

where N_{total} is the total number of single game outcomes that were simulated. From this, we have a numerical estimate of the probability distribution $P(i|X0) = P(X|H0)$ (single outcome) with the following equation.

$$P(i|H0) = \frac{N_i}{N_{total}} \tag{6}$$

These histograms of the vectors of data are then plotted as histogram to help visualize the data for each hypothesis. This plot and its organization provide the framework to numerically estimate the probability distribution for each hypothesis. These plots are titled, "Histogram of Simulated Data" and an example of this plot is shown in Figure 4.

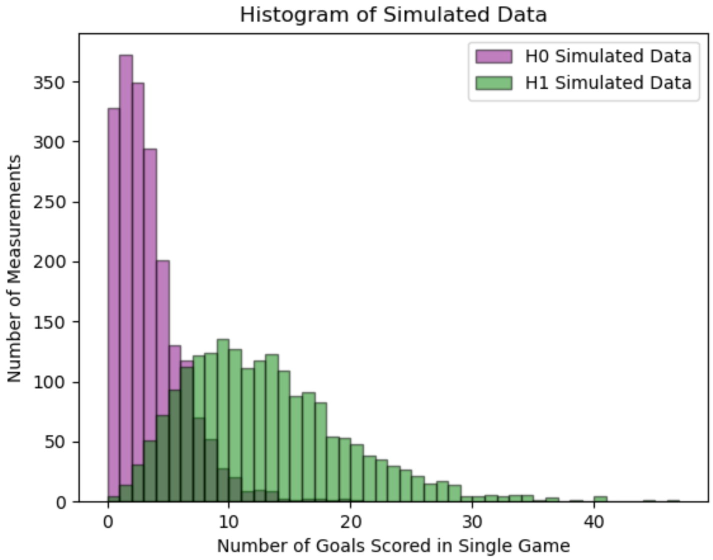

Figure 4: A plot of the vectors of outcome data shown in a histogram with integer binning. This plot and its organization provide the framework to numerically estimate the probability distribution for each hypothesis.

In addition to the histogram plot of the data, these functions of the estimated probability distribution were also plotted. This function details the probability that a certain measurement (i.e. number of goals scored in a game) came from the scenario described by each hypothesis. These functions are shown in Figure 5.

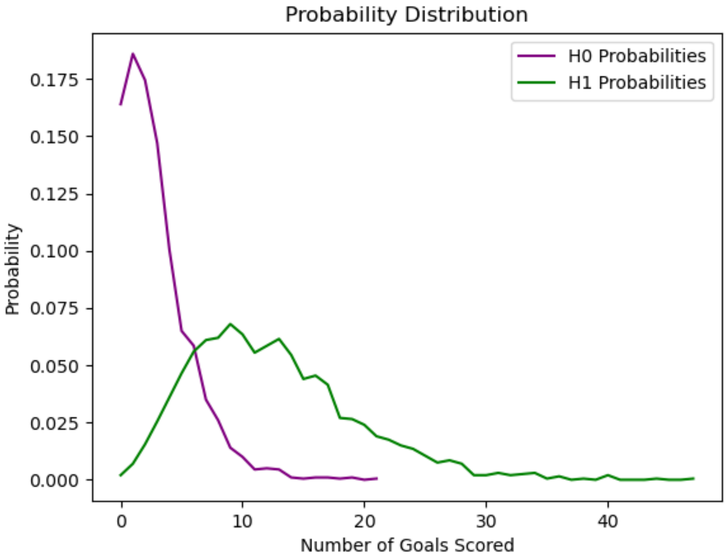

Figure 5: A plot showing the estimated probability distribution for each hypothesis distribution. From this plot, you can discern how likely a given measurement is to come from the different hypotheses.

This probability function additionally helps to construct our likelihood calculations as we as the analogous function for the alternative hypothesis $H1$. With these numerical functions in hand, $P(i|H0)$ and $P(i|H1)$, we can now calculate the log likelihood ratios. In the experimental setup, we simulated many "experiments." Each experiment is made up of N_{meas} games with a single outcome measurement (e.g. 5 goals scores, 8 goals scored, etc.). The simulation is made up of N_{exp} number of different cases of these N_{meas}. Thus, due to this setup, the log likelihood ratio (Λ or LLR) for a single experiment of N_{meas} measurements is as follows:

$$\Lambda = \sum_j [log(P(X_j\ H1)) - log(P(X_j\ H0)] \tag{7}$$

where j is indexing the number of measurements (N_{meas}) from the experiment. In other words, each X_j is the number of goals scored in the j-th game out of N_{meas} measurements.

This log likelihood ratio is calculated for each experiment in the simulation and stored in an array which will be visualized as a histogram shortly.

One caveat that should be mentioned for this algorithm is a minor, numerical "band-aid" that had to be applied so that the code would run correctly. When considering different hypotheses, we ran into the case that one hypothesis "never" gave a value that appeared in the other hypothesis' data (at least not within the number of runs we did). In this case, the probability of getting that value would be zero and thus cause a discontinuity in our log likelihood function. This could potentially be a numerical shortcoming as the value may just be very unlikely, but could also be a symptom that is indicative that the two hypotheses that are being testing are too dissimilar. To account for this, we assigned a "minimum" probability for values that were never observed in the simulation for one hypothesis that should still be possible. Thus, the probability that is assigned to such values is a probability of $\frac{1}{N_{sim}}$ where N_{sim} is the total number of outcomes observed. This is not necessarily correct or careful, but if N_{sim} is large, then it is effectively an upper-bound on a very small probability.

After sorting has concluded, the hypothesis testing can finally begin utilizing these arrays of the log likelihood ratios for each hypothesis. The first step in the process for this next stage is to define a significance level α for our test. In this case, the values of $\alpha = 0.02$ was chosen which corresponds to the test being accurate 98% of the time (a confidence level of 0.98 or $(1 - \alpha)$). This sets our false positive rate also known as a Type 1 error. This means that we will tolerate a false positive rate 2% of the time. The α parameter is also chosen ahead of time in an effort to avoid experimental bias. A visual representation of this parameter is shown in Figure 6

Using this value for α, we will find a test statistic value in the array for H_0 which corresponds to 98% of the area under the curve to be one one side and 2% of the area being on the left. Essentially, we find the quantile for 0.98. This value is known as the critical test statistic and this value for the data we have simulated is sown in Figure 7 as a red line. The next step is to define the value for β which as can be seen from Figure 6 is the area for the other distribution (in this case H_1) that is sectioned off by the critical Λ value. This is known as a Type II error which corresponds to our false negative rate. For this case, the value for beta (β) was determined to be 0.005 which means that the line did not section off much data for the H_0 distribution. The value for β helps us to quantify the Power of the Test which is calculated simply with the expression $1 - \beta$. In this case, the power of the test is 0.995. This value of β shows a very powerful test since the false negative rate is very low. The most powerful test would correspond to a value of 1.

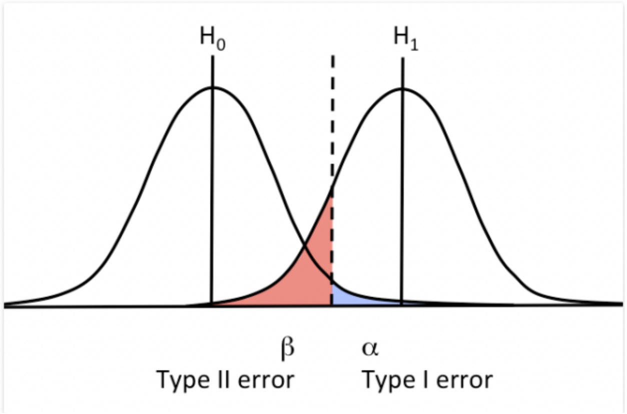

Figure 6: Figure depicting how α and β are determined from a graph. The dashed line would correspond to the critical Λ in this case. [7]

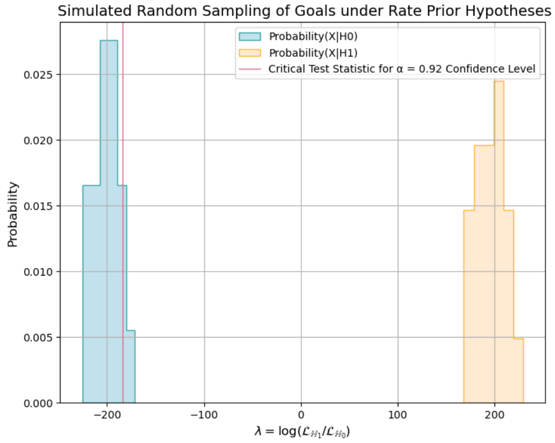

Figure 7: Figure depicting the LLRs for both hypotheses. The red line corresponds to the portion of the distribution that corresponds to the critical test statistic. [7]

This concludes the formal hypothesis testing that the code conducts and now the results can be interpreted under the scenario that we crafted. The following section 5 will show the results for the two sets of hypotheses previously described.

5 Hypothesis Testing Discussion

Now that the hypotheses, data generation, and analysis have been fully explained. The results that come from testing the two sets of hypotheses are shown in the following two subsections Sections 5.1 and 5.2. For each situation, all four plots described earlier are shown.

5.1 Separated Hypotheses

These hypotheses demonstrate what happens in the analysis when the hypotheses turn out to be too dissimilar to correctly test these hypotheses to against each other. In essence, it means that they are too easy to tell apart that hypothesis testing is not necessary. The following Figures 8, 9, 10, and 11 show these results.

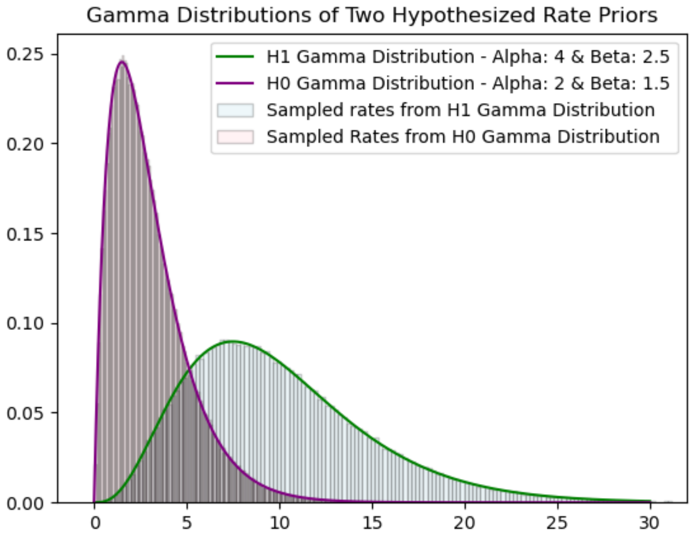

Figure 8: A comparison of the gamma distribution that results from the user inputted α and β parameters along with a histogram showing all of the values that were sampled from the gamma distribution over the course of the experiment.

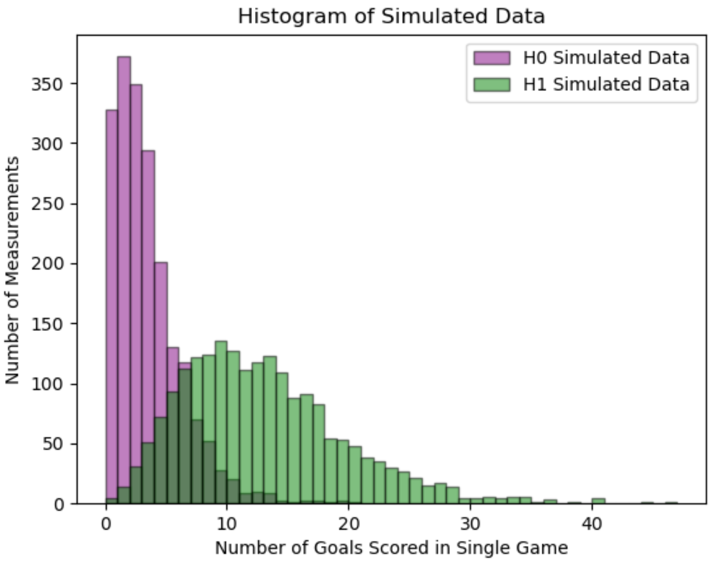

Figure 9: A plot of the vectors of outcome data shown in a histogram with integer binning. This plot and its organization provide the framework to numerically estimate the probability distribution for each hypothesis.

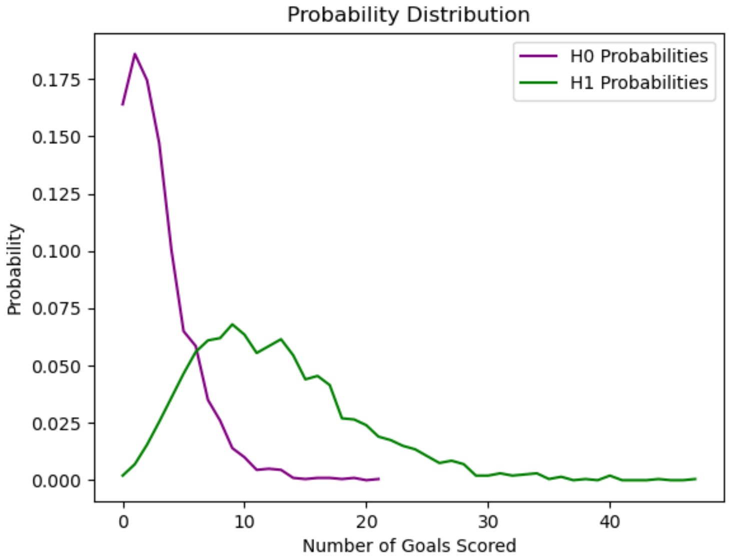

Figure 10: A plot showing the estimated probability distribution for each hypothesis distribution. From this plot, you can discern how likely a given measurement is to come from the different hypotheses.

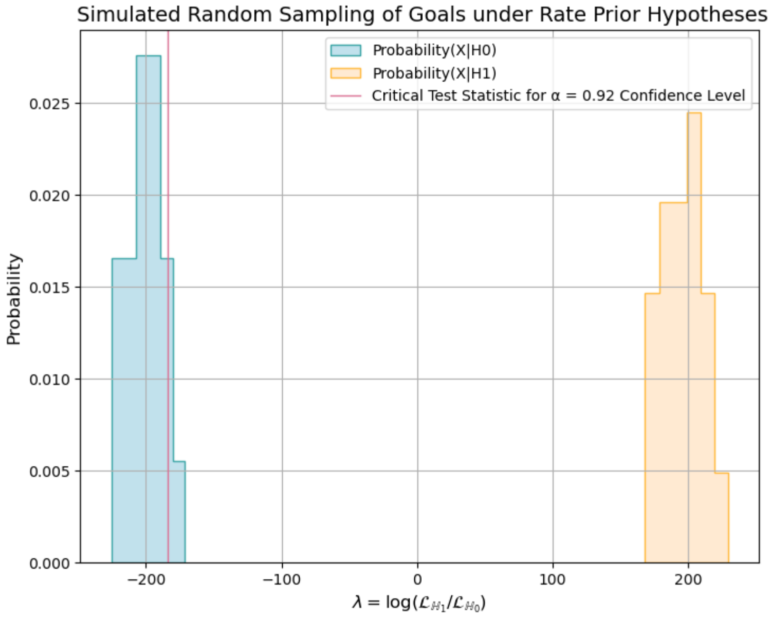

Figure 11: Figure depicting the LLRs for both hypotheses. The red line corresponds to the portion of the distribution that corresponds to the critical test statistic. [7]

5.2 Overlapped Hypotheses

Now, for the second set of hypotheses, we can see what the output is when the hypotheses are similar enough to be hypothesis tested. These hypotheses have prior gamma distribution which are much more similar and should yield log likelihood ratio plots that overlap so that hypothesis testing can be performed. First the gamma distributions along with the sampled rates are shown below in Figures 12 and 13.

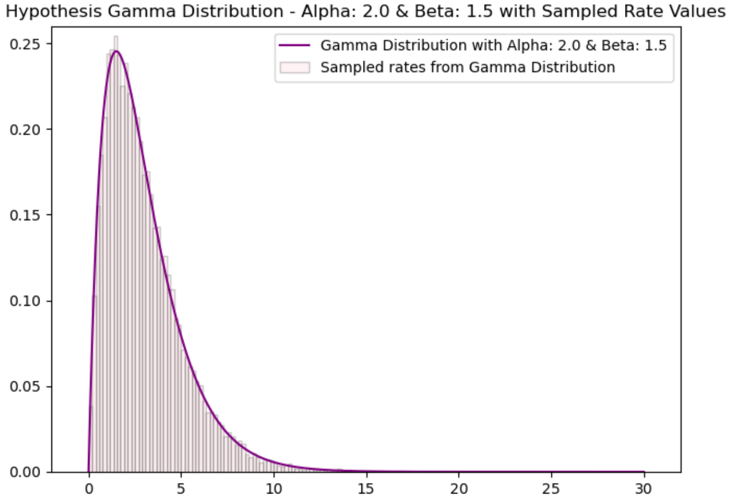

Figure 12: A comparison of the gamma distribution that results from the user inputted $\alpha = 2.0$ and $\beta = 1.5$ parameters along with a histogram showing all of the values that were sampled from the gamma distribution over the course of the experiment.

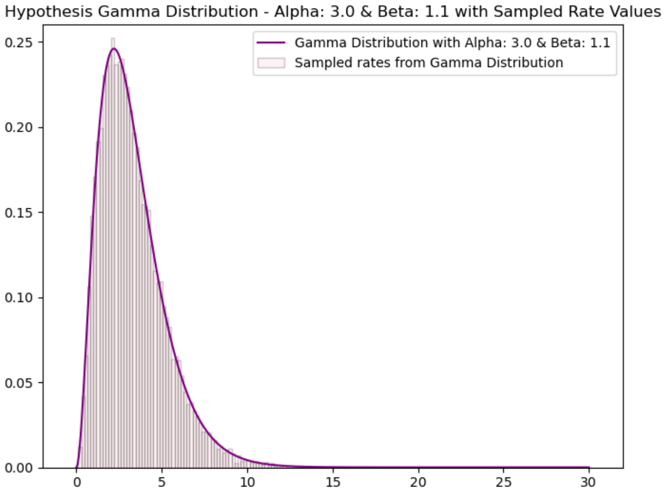

Figure 13: A comparison of the gamma distribution that results from the user inputted $\alpha = 3.0$ and $\beta = 1.1$ parameters along with a histogram showing all of the values that were sampled from the gamma distribution over the course of the experiment.

With these two prior distributions, the rest of the analysis was conducted which demonstrated that they were indeed similar enough to be hypothesis tested. First, in Figure 14 it can be seen that the simulated data is much more similar than that of Figure 9. Additionally, in Figure 15 the numerical probability distribution estimates are also shown. These show just how intertwined the probability distributions are. It is less clear in this case, as compared to the previous case shown in Figure 10. This is not one clear region that H0 is more likely and one clear region where H1 is more likely. This is a good indication that they are similar enough to yield good hypothesis testing results and is also a good indication that our analysis can determine meaningful results from hypotheses that are only slightly different.

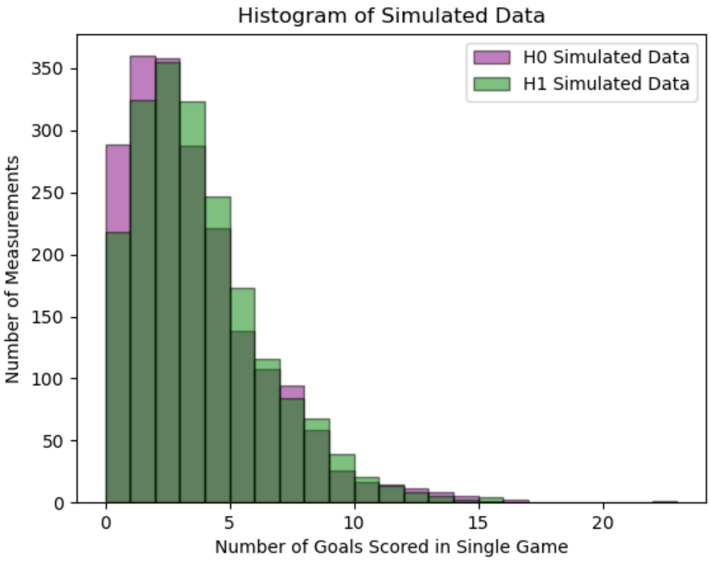

Figure 14: A plot of the vectors of outcome data shown in a histogram with integer binning. This plot and its organization provide the framework to numerically estimate the probability distribution for each hypothesis.

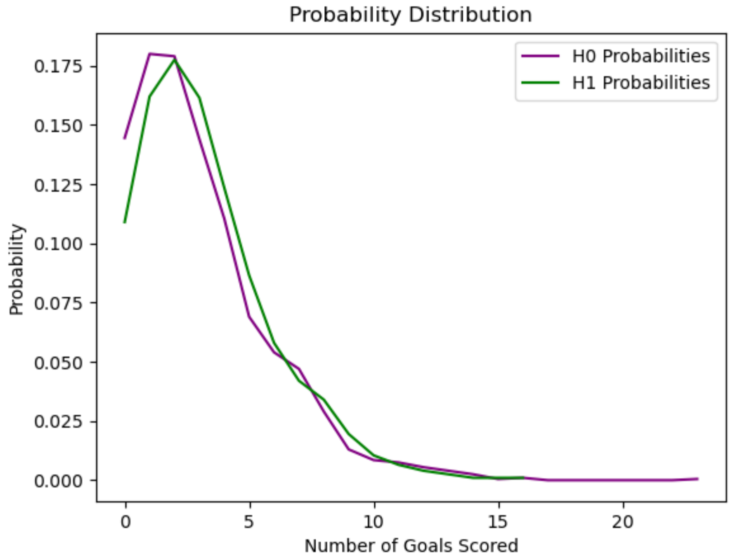

Figure 15: A plot showing the estimated probability distribution for each hypothesis distribution. From this plot, you can discern how likely a given measurement is to come from the different hypotheses.

Finally, the log likelihood ratio histogram can be plotted. As shown in Figure 16. In this case, we can see that the LLR histograms do in fact overlap, albeit slightly. The red line indicates where the critical lambda (Λ) falls for the H0 hypothesis and H1 hypothesis. The Λ had a value of 0.42055 in this case in the H0 hypothesis. When compared to where this fell in the H1 distribution. This yielded beta value of $\beta = 0.018405$ and, subsequently, a Power of the Test of 0.981595. This is indicative of a very strong test in this case especially given the alpha value that was used ($\alpha = 0.98$). Up until this point, we have only considered our hypotheses with rather arbitrary decisions for the prior gamma distribution. In this project, we set the parameters of alpha and beta simply to resemble different distributions to test. In the case of a more real-world sample, one could assign real values to those parameters and could even add in more nuisance parameters in a more complicated model. Additionally, it would be a good test to see what the potential minimum number of results is needed to achieve a certain level of confidence. This could help set a limit ot see if a sample of real-world data could be used to make conclusions about the team's performance that are meaningful.

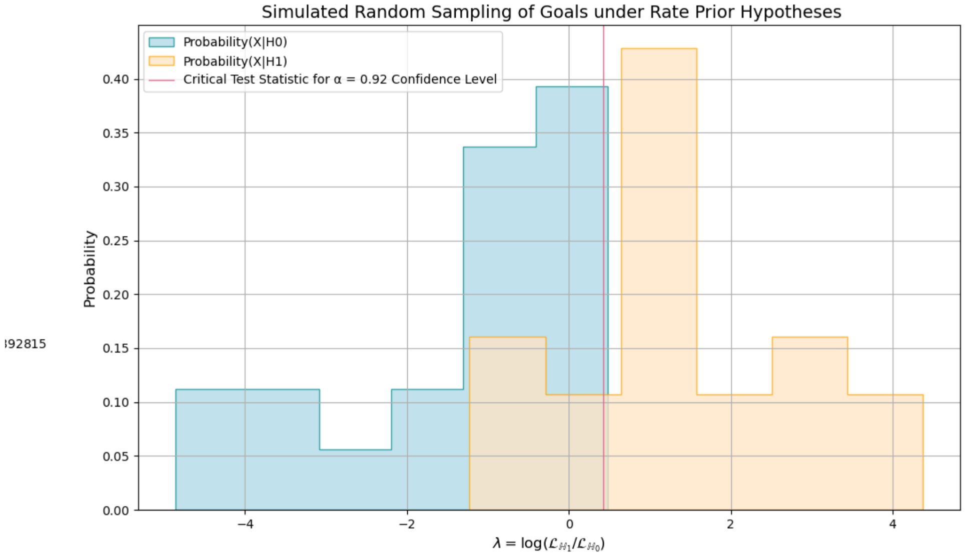

Figure 16: Figure depicting the LLRs for both hypotheses. The red line corresponds to the portion of the distribution that corresponds to the critical test statistic. [7]

6 Conclusion

Overall, this simulated experiment demonstrates how data for two competing sets of hypotheses can be randomly sampled according to a Poisson-Gamma model where we employ the gamma distribution as a prior distribution to the Poisson. The simulated data that resulted from this model could then be interpreted and analyzed by calculating the numerical probability distribution, log likelihood ratio, and finally the performance of hypothesis testing with two different sets of hypotheses. The first set considered the situation where two hypotheses were chosen which were too dissimilar to perform hypothesis testing (i.e. it was too easy to tell the two hypotheses apart). Alternatively, the second set of hypotheses showed the case of two hypotheses that were similar enough to be tested and compared. In this case, it was determined that the hypotheses could be distinguished with a rather high power of the test. This means that the analysis methods that were utilized are able to meaningfully distinguish between two hypotheses even if they are very similar. Even without the real world data set present, we are able to quantify the separation of the distributions through the value of β and the power of the test under a certain significance level α. So next time you tune in to watch your favorite soccer team compete or are lucky enough to watch in person, think about taking goal data to apply this method in quantifying their performance for the season and ponder what other factors would be useful to consider in making a simulation model!

7 Repository Link

GitHub Repository Link: `https://github.com/aelieber1/PHSX815_Project2`

References

[1] T. Tolonen and V. Hyvönen, *Bayesian inference 2019*, Mar, 2019.

[2] J. Frost, *Gamma distribution: Uses, parameters amp; examples*, Aug, 2022.

[3] VrcAcademy, *Gamma distribution calculator*, May, 2021.

[4] A. Kumar, *Poisson distribution explained with python examples*, Oct, 2021.

[5] J. Frost, *Poisson distribution: Definition amp; Uses*, May, 2022.

[6] O. Eaton, *Modelling the Distribution of Football Goals*,.

[7] D. Darrin and D. Darrin, *Type I and type II error*, Oct, 2022.