Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6015: Introduction to Enterprise Analytics**

**Assignment:**

Module 1 Project-  Analysis of a Betting Strategy in Sports

**Submitted on:**

February 28, 2023

**Submitted to:**  **Submitted by:**

Prof: AZADEH MOBASHER NIKSHITA RANGANATHAN

# **ABSTRACT**

Probability theory is a subfield of mathematics that is concerned with the study of random events or phenomena. It provides a framework for quantifying and analyzing the likelihood of an event occurring based on certain assumptions or conditions. It is a fundamental tool for comprehending and modeling uncertainty, making predictions, and making informed decisions.

Probability theory has numerous practical applications in diverse domains, including statistics, finance, engineering, physics, and computer science. In finance, for example, probability theory is used to model and analyze the behavior of financial assets and markets, while in engineering, it is used to assess the reliability and safety of systems and structures, and in computer science, it is used to develop algorithms for machine learning and data analysis.

Continuous and discrete probability distributions are two types of mathematical models used to describe the probability distribution of random variables. The main difference between the two is that a continuous distribution has an infinite number of possible values within a given range, while a discrete distribution has a finite or countably infinite number of possible values.

Common types of continuous probability distributions:

* Normal distribution, also known as the Gaussian distribution, is characterized by a bell-shaped curve and is used to model many natural phenomena, such as human height or IQ scores.
* Uniform distribution, which is frequently used in simulations and random number generation, is defined by a constant probability density function over a particular range.
* Exponential distribution is used to model the time between events in a Poisson process, where events occur randomly and independently over time.

Common types of discrete probability distributions:

* A binary result, such as a coin toss, where the chance of success is p and the probability of failure is 1-p, is modeled using the Bernoulli distribution.
* The binomial distribution is used to model the number of successes in a given number of independent Bernoulli trials.
* Poisson distribution is used to model the frequency of occurrences of an event over a defined period of time, such as the number of customers entering a store in a given hour.

**Chi-Square test:**

The Chi-square test is an approach used to evaluate if there is a difference between projected and actual frequencies between various groups. It is used to evaluate the independence between two events or to see if there is a significant correlation between two variables.

The advantages of this test are that it is relatively easy to calculate and interpret and can be used to compare large numbers of variables. Additionally, it is a non-parametric test, meaning it does not require any assumptions about the distribution of the data.

The Chi-square test is utilized in numerous applications. In marketing research, it is used to compare the efficacy of multiple advertising campaigns, and in medical research to evaluate the effectiveness of different treatments. Additionally, it is used in areas like survey analysis, quality assurance, and genetics. In a variety of industries, it is a useful tool for data analysis and decision-making.

**INTRODUCTION**

Millions of Americans tune in to watch their favorite teams play in baseball, making it one of the most watched sports in the country. One of the most iconic rivalries in all of sports is the rivalry between the Boston Red Sox and the New York Yankees, two of the most successful teams in Major League Baseball.

Probability theory has applications in various fields, and one such field is sports betting. In this project, we will be applying probability theory to evaluate the betting strategies in a series of baseball games between Boston's Red Sox and New York's Yankees.

We will calculate the probability of the Red Sox winning the series in three different scenarios:

* In the first scenario, we will analyze a best-of-three series between Boston Red Sox and New York Yankees, with the first game being played in Boston, the second game being played in New York, and the third game being played in Boston again.
* In the second situation, the series is a best-of-three format with games 1 and 3 taking place in New York, and 2nd game in Boston.
* In the third scenario, a best-of-five series will be played, with games alternating between Boston and New York and the first game being played in Boston. The winner will be the first squad to win three games.

It is mentioned that the Red Sox have a 0.6 probability of winning a game at their home stadium, while the Yankees have a 0.57 probability of winning a game at their home stadium. For each game, we will win $500 if the Red Sox win and lose $520 if they lose the game.

The outcome of each game is assumed to be independent of each other, meaning that the result of one game does not affect the results of the other games.

For each scenario, we will calculate the probability of winning the series and the expected net win. We will also generate 10,000 random values to estimate the expected net win with a 95% confidence interval and create a frequency distribution to compare the observed and theoretical frequencies using the Chi-squared goodness of fit test

**ANALYSIS & INTERPRETATION**

**Part – 1 (First game in Boston, Second game in New York, Third game in Boston)**

**(i) Calculate the probability that the Red Sox will win the series.**

**Table

Description automatically generated**

**Figure 1- Table with net probabilities of 6 cases (Part-1)**

Based on the above table, we can see that the Red Sox team has **six possible cases** against the New York team, and each case has a different probability of winning or losing a game. The net probability column shows the overall probability of each case occurring, calculated as the product of the individual probabilities.

The net probabilities of the cases are added up to get the likelihood of winning or losing the series.

Red Sox winning the series is possible in **Case 1 (Win, Win), Case 2(Win, Lose, Win), and Case 3 (Lose, Win, Win).**

Adding these probabilities, we get:

**0.26 + 0.21 + 0.10 = 0.57**

The probability of the Red Sox winning the series is **0.57** or **57%.**

**(ii) Create a probability distribution for your net win series (X), and then determine its mean and standard deviation.**

**Table

Description automatically generated**

**Figure 2- Net wins (Part-1)**

In figure 2, we can see the bets for each game, as well as the resulting net wins, which shows the total amount of money depending on the outcomes of the games and the bets placed.

**Table

Description automatically generated**

**Figure 3- Probability distribution of Net wins (Part-1)**

This table shows the probability distribution of a random variable (Net wins). There can be 4 possible values of Net wins.

Table

Description automatically generated

**Figure 4- Expected value, Variance, and Standard deviation (Part-1)**

The expected value, variance, and standard deviation are statistical measures used to describe the distribution of values in a dataset. They are used to explain and explore the behavior of variables, as well as to draw conclusions and predictions from statistics using probability distributions.

The standard deviation is **795.15**, and the expected Mean value is **57.89**.

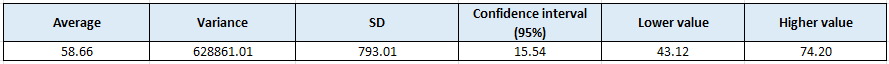
**(iii) Create 10,000 random values for X. Calculate a 95% confidence interval using the Y values to estimate the expected net win. Does this confidence interval contain E(X)?**

**Table

Description automatically generated**

**Figure 5- 10,000 random values (Part-1)**

A series of 10,000 random values for net wins were generated using the RAND() function in Excel.



**Figure 6- Statistics of Y (Part-1)**

The statistics of Y, including the mean (average) and standard deviation, have been calculated using Excel functions, and a confidence interval has been computed using the CONFIDENCE() function.

The average of the generated numbers is positive, as indicated by the mean value of Y, which is **58.66**. The standard deviation is **793.01**. The confidence interval, calculated with a 95% confidence level, has a margin of error of **15.54**.

This results in a lower limit of **43.12** and an upper limit of **74.2** for the confidence interval, indicating the range of values within which the true population mean is expected to fall with a 95% confidence.

**(iv) Construct a frequency distribution for Y. Next, use the Chi-squared goodness of fit test to verify how closely the distribution of Y has estimated the distribution of X.**

**Table

Description automatically generated**

**Figure 7- Theoretical and observed frequencies (Part-1)**

The Theoretical Frequency column denotes the expected frequency of each value based on the given probabilities, while the Observed Frequency column shows the actual frequency of each value in the observed dataset

Chi-Squared measures the goodness of fit by comparing observed and expected frequencies for each value, then summing the squared differences divided by the theoretical frequency. Smaller values indicate a better fit.

Shape

Description automatically generated with low confidence

**Figure 8- Chi-Square goodness of fit test (Part-1)**

* In this case, the Chi-Squared Total value is **3.29**, which suggests a relatively small deviation between the observed data and the expected distribution. Degrees of freedom is **3**.
* The null hypothesis of a Chi-Squared test is that there is no significant variance between the observed data and the expected distribution.
* Alternate hypothesis is that there is a difference between observed and theoretical frequency.
* Alpha value is **0.05**. In this case, the P-value is **0.35**, which is greater than the alpha value.
* We can infer that there is insufficient evidence to reject the null hypothesis. In other words, the observed data and expected distribution do not deviate significantly from one another.

**(v) Describe whether your betting strategy is favorable to you**

Based on the analysis, it appears that the betting strategy is favorable for me, as the generated values indicate a positive average. Also, the mean value of my total earnings is expected to be between **43.12** and **74.2** with 95% confidence. These values indicate that I am likely to make a profit if I place bets using this strategy.

**Part – 2 (First game in New York, Second game in Boston, Third game in New York)**

**(i) Calculate the probability that the Red Sox will win the series.**

**Table

Description automatically generated**

**Figure 9- Table with net probabilities of 6 cases (Part-2)**

Again, we will have **6 cases** for Part 2 as well which are described by the outcomes of each game, such as win or lose for each team, and the corresponding probabilities of those outcomes.

These are the three ways in which Boston can win the series:

* Win two consecutive games
* Win the first and the third game but lose the second game
* Lose the first game but win the next two games.

The probability of the Red Sox winning the series is **0.48** and is calculated by :

**0.26 + 0.07 + 0.15 = 0.48**

**(ii) Create a probability distribution for your net win series (X), and then determine its mean and standard deviation.**

**Table

Description automatically generated**

**Figure 10- Net wins (Part-2)**

Net Wins (X) values are **1000**, **480**, **-1040**, and **-540** which is the same as the Part-1 of the assignment.

**Table

Description automatically generated**

**Figure 11- Probability distribution of Net wins (Part-2) with E(X), Variance and Standard deviation**

The expected value is the average value that we would expect to get if we repeated the experiment many times. In this case, the expected value is **- 31.24.**

The variance and standard deviation of the distribution provide information about the spread or variability of the outcomes around the expected value. The variance of X is **639,984.86** and the standard deviation is **799.99**. The standard deviation is relatively large, indicating that the distribution is spread out.

**(iii) Create 10,000 random values for X. Calculate a 95% confidence interval using the Y values to estimate the expected net win. Does this confidence interval contain E(X)?**

**Table

Description automatically generated**

**Figure 12- 10,000 random values (Part-2)**

****

**Figure 13- Statistics of Y (Part-2)**

The average Y is estimated to be **-37.26**, with a variance of **641292.38** and a standard deviation of **800.81**. The 95% confidence interval for the expected Y is **(-52.96, -21.56)**, which means that we can be 95% confident that the true expected Y falls within this range.

**(iv) Construct a frequency distribution for Y. Next, use the Chi-squared goodness of fit test to verify how closely the distribution of Y has estimated the distribution of X**

**Table

Description automatically generated**

**Figure 14- Theoretical and observed frequencies (Part-2)**

* The expected or theoretical frequency distribution is calculated based on the probability distribution of the random variable, in this case, Net Wins (X)



**Figure 15- Chi-Square goodness of fit test (Part-2)**

* **NULL Hypothesis**: This is the default hypothesis that there is no significant difference between the theoretical frequency distribution and the observed frequency distribution.
* **ALTERNATE Hypothesis**: Contrary to the NULL hypothesis, the hypothesis here is that there is a significant difference between the theory and observed frequency distributions.
* The level of significance is set to **0.05** (95% confidence). This means that if the P-value obtained is less than or equal to 0.05, the null hypothesis can be rejected at the 95% confidence level.
* In this case, since the p-value is high **(0.85)**, we fail to reject the null hypothesis and conclude that the observed frequencies are consistent with the theoretical distribution.

**(v) Describe whether your betting strategy is favorable to you**

Based on the observations, it has been determined that the betting strategy is not good. The 95% confidence interval of the mean value of their total earnings falls between **-52.96** and **-21.56**. Since both of these values are negative, it indicates that the person is likely to lose money if they continue to place bets using this particular betting strategy

**Part – 3 (Best of five series)**

**(i) Calculate the probability that the Red Sox will win the series.**

**Table

Description automatically generated**

**Figure 16- Table with net probabilities of 20 cases (Part-3)**

Unlike, Part 1 and 2, we have 20 scenarios in Part 3.

Overall, the table suggests that the Red Sox have a higher probability of winning the series than the Yankees, with a total probability of **0.56** for the Red Sox to win and **0.44** for the Yankees to win.

**(ii) Create a probability distribution for your net win series (X), and then determine its mean and standard deviation.**

**Table

Description automatically generated**

**Figure 17- Net wins (Part-3)**

The X values here are **1500**, **980**, **460**, **-560**, **-1060,** and **-1560**.

**Graphical user interface

Description automatically generated**

**Figure 18- Probability distribution of Net wins (Part-3) with E(X), Variance and Standard deviation**

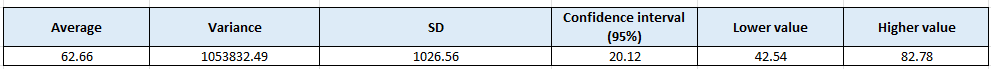
The expected net win in a series of bets is **$76.35** and the standard deviation is **1,024.73**. These values give us important information about the distribution of the net win and can help us make informed decisions about betting strategies.

**(iii) Create 10,000 random values for X. Calculate a 95% confidence interval using the Y values to estimate the expected net win. Does this confidence interval contain E(X)?**

**Table

Description automatically generated**

**Figure 19- 10,000 random values (Part-3)**

****

**Figure 20- Statistics of Y (Part-3)**

The average (or expected value) of the Y is **62.66**, with a standard deviation of **1026.56**. The lower and higher values of the confidence interval are **42.54** and **82.78**, respectively. This means that we can be 95% confident that the true mean of Y in the series lies between **$42.57** and **$82.78**.

**(iv) Construct a frequency distribution for Y. Next, use the Chi-squared goodness of fit test to verify how closely the distribution of Y has estimated the distribution of X**

**Table

Description automatically generated**

**Figure 21- Theoretical and observed frequencies (Part-3)**

**Figure 22- Chi-Square goodness of fit test (Part-3)**

* In this case, there are 6 categories, so the degree of freedom is 5.
* The p-value is **0.30**, which is greater than the commonly used threshold of 0.05. Therefore, we would fail to reject the null hypothesis that there is no association between the two and conclude that suggests that there is no significant evidence to say that the observed frequencies match the expected frequencies.

**(v) Describe whether your betting strategy is favorable to you**

According to the observed data, it can be concluded that this betting strategy is profitable. With 95% confidence, it can be stated that my expected net earnings would be between **42.54** and **82.78**, both of which are positive values indicating that I can expect to make money by placing bets on this series.

**CONCLUSION**

This assignment has provided insight into the probability of winning a series of bets and has helped to determine the best betting strategy. This has been done by calculating the probability of winning and constructing probability distributions, as well as calculating confidence intervals and performing Chi-squared tests.

Part 1 of the assignment showed that there was a 57% chance of the Red Sox winning the series if they played the first game in Boston and the second and third games in New York. The expected mean and standard deviation of the net win series were 57.89 and 795.15, respectively. The confidence interval contained the expected value, indicating that the betting strategy is favorable.

The second strategy, from Part 2, has a probability of 48% of the Red Sox winning the series, which is lower than the strategy from Part 1. The expected value is -31.24, with a variance of 639,984.86 and a standard deviation of 799.99. The 95% confidence interval indicates that the expected net win is likely to be between -52.96 and -21.56. This strategy is not as favorable as the one from Part 1.

In Part 3, the probability of the Red Sox winning the series was 56%, and the expected value was 76.35 with a standard deviation of 1,024.73.

The analysis in this assignment has shown that these betting strategies can be beneficial or detrimental to the bettor and that, understanding the probability of the outcomes and their expected net win can help the bettor make informed decisions about their betting strategy.

**REFERENCES**

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

Home - RDocumentation. (n.d.). <https://www.rdocumentation.org/>