Module 1

ALY 6050: Introduction to enterprise Analytics

**Project: Analysis of a Betting Strategy in Sports**



**Submitted to:**

Prof. Roy Wada

**Submitted by:**

Vartika Joshi

(NUID: 002760320)

**INTRODUCTION**

In this assignment, problem involving the Boston Red Sox and New York Yankees playing a best of three series, where the winner is the first team to win two games. The probability that each team would win in their home stadium, as well as the potential payout for betting on each game. In parts 1-3, assuming the outcomes were independent, The task is with calculating the probability of the Red Sox winning the series, constructing a probability distribution for my net win (X), and estimating expected net win through random values (Y) and a confidence interval. Additionally, to create a frequency distribution for Y and use the Chi-squared goodness of fit test to verify how closely the distribution of Y estimated the distribution of X.

**PART 1-**

The Boston Red Sox is a professional baseball team that is scheduled to play a three-game series. To calculate the probability of the team winning the series, the possible outcomes of the series is considered. This report presents an analysis of the probabilities of the possible outcomes and the overall probability of the team winning the series.

1. To calculate the probability that the Red Sox will win the series, I considered all the possible outcomes of the three-game series. The team can win the series by winning the first two games, winning the first and third games, or winning the second and third games. The probabilities of these outcomes are as follows:

Red Sox win first two games: (0.59)^2 = 0.3481

Red Sox win first and third games: 0.59 \* (1 - 0.55) \* 0.59 + (1 - 0.59) \* 0.55 \* 0.55 = 0.3128

Red Sox win second and third games: (1 - 0.59) \* 0.55 \* 0.59 = 0.1595

So the probability that the Red Sox will win the series is the sum of these probabilities: 0.3481 + 0.3128 + 0.1595 = 0.8204 or approximately 82.04%.

1. To construct the probability distribution, I consider the four possible outcomes and their corresponding probabilities and net win/loss.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Outcome** | **Probability** | **Net Win/Loss($)** |
| Red Sox win in 2 games | 0.3481 | 1010 |
| Red Sox win in 3 games | 0.3128 | 485 |
| Yankees win in 3 games | 0.0857 | -545 |
| Yankees win in 2 games | 0.1684 | -1050 |

Table 1: Outcome and the probabilities of 4 scenarios

From this information, the probability distribution for X can be constructed as follows:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **X** | **Net Win/Loss** | **P(X=x)** |
| $1,010 | Positive | 0.3481 |
| ($545) | Negative | 0.3128 |
| ($1,050) | Negative | 0.2541 |

Table 2: Probability distribution of X

Expected Value, Variance, and Standard Deviation: To calculate the expected value, I take the sum of the products of each possible net win/loss and its corresponding probability. Therefore,

E(X) = $1010 \* 0.3481 + (-$545) \* 0.3128 + (-$1050) \* 0.2541 = -$63.28

This means that, on average, a person playing this series would expect to lose $63.28.

To calculate the variance, I first find the deviation of each possible net win/loss from the expected value, square it, and multiply it by the corresponding probability. I then sum up these products to get the variance. Therefore,

Var(X) = $648070.70

This means that the variation in net win/loss is quite high and that players can expect to experience a wide range of outcomes.

To calculate the standard deviation, we take the square root of the variance. Therefore

SD(X) = sqrt(Var(X)) = $805.5

This means that, on average, a person playing this series can expect to lose $67.28 with a standard deviation of $805.5.

1. I created 10,000 random values for X in Excel using the RAND() function and assigned each random number to a possible outcome of the series based on the probabilities provided in part (ii).

Using the generated random values for X, I estimated the expected net win using a 95% confidence interval. To do this, I used the AVERAGE() and STDEV() functions in Excel to calculate the sample mean and standard deviation of the 10,000 random values for X. The sample mean was found to be 53.1315, while the variance was 649204.6967 and the standard deviation was 805.7323977.

The 95% confidence interval for the expected net win was calculated, I found that the confidence value (95%) was 15.79. The lower limit of the confidence interval was 37.34 and the upper limit was 68.92.

Based on the generated random values for X and the confidence interval calculated, I can conclude that the estimate of the expected net win is consistent with the true value.

1. To create a frequency distribution for Y, I grouped the values into discrete intervals using a bin size of $50. I selected the range of cells containing the random values for Y.

Chi-Squared Goodness of Fit Test:

To perform the chi-squared goodness of fit test, I created a table with two columns, one for the observed frequency of each bin in the histogram and one for the expected frequency of each bin based on the probability distribution of X from part (ii). I calculated the expected frequency for each bin by multiplying the total number of observations (10,000) by the probability of X falling in that bin. I then calculated the chi-squared statistic by summing the squared difference between the observed frequency and expected frequency for each bin, divided by the expected frequency.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Theoretical Frequency** | **Observed Frequency** | **Chi-Squared** |
| 2655 | 2628 | 0.27 |
| 3003.1 | 3062 | 1.16 |
| 2086.9 | 2029 | 1.61 |
| 2255 | 2281 | 0.30 |

Table 3: Chi-Squared Goodness of Fit Test

The degrees of freedom for the chi-squared test are calculated by subtracting 1 from the number of bins, which gives us 3 degrees of freedom in this case. The p-value for the chi-squared test is calculated using the CHISQ.DIST.RT function in Excel.

The p-value of the chi-squared test is 0.656, which is greater than the significance level of 0.05. Therefore, **we fail to reject the null hypothesis and conclude that the observed frequency distribution of Y** is consistent with the expected probability distribution of X. In conclusion, the chi-squared goodness of fit test indicates that the observed frequency distribution of Y does not significantly differ from the expected probability distribution of X.

1. After performing the analysis, the results indicate that the null hypothesis will not be rejected based on the P-value. The P-value is greater than 0.05, which means that we cannot reject the null hypothesis. Therefore, we cannot conclude that our betting strategy is favorable, and we cannot expect to make a profit in the long run.

**PART 2-**

(i) To calculate the probability that the Red Sox will win the series, I created a table in Excel with the outcomes and probabilities for each possible scenario:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Probability of Boston Red Sox winning at their home stadium | | 0.59 |
| Probability of Boston Red Sox losing at their home stadium | | 0.41 |
| Probability of New York Yankees winning at their home stadium | | 0.55 |
| Probability of New York Yankees losing at their home stadium | | 0.45 |

Table 4: Probabilities for each team to win in a specific location.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | **New York** | **Boston** | **New York** | **Product of Probability** |
| Boston Red Sox wins the series | 0.45 | 0.59 | - | 0.27 |
| 0.45 | 0.41 | 0.45 | 0.08 |
| 0.55 | 0.59 | 0.45 | 0.15 |
| New York Yankees wins the series | 0.55 | 0.41 | - | 0.23 |
| 0.55 | 0.59 | 0.55 | 0.18 |
| 0.45 | 0.41 | 0.55 | 0.10 |

Table 5: Probabilities of each team winning the series.

Based on the calculations on excel, the probability of Boston Red Sox winning the series is 0.49455 or approximately 49.5%, while the probability of Boston Red Sox losing the series is 0.50545 or approximately 50.5%.

These probabilities were calculated by considering the various outcomes of the series based on the probabilities of Boston Red Sox winning or losing at their home stadium and the probability of New York Yankees winning or losing at their home stadium.

(ii) I used Excel to create a table to organize the information and calculate the probability distribution, expected net win, and standard deviation of the net win. First, I created a table to represent the possible outcomes of the three games, based on which team wins each game.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| **Game 1** | **Game 2** | **Game 3** | **Net Win** |
| Yankees | Red Sox | Yankees | ($525) |
| Red Sox | Yankees | Yankees | ($525) |
| Yankees | Red Sox | Red Sox | $505 |
| Red Sox | Yankees | Red Sox | $505 |
| Yankees | Yankees | N/A | ($525) |
| Red Sox | Red Sox | N/A | $505 |

Table 6: Possible outcomes of three games

Then created a probability distribution table and calculated the expected net win and standard deviation using Excel formulas. Based on the given probabilities the expected net win is -$36.75 with a standard deviation of $472.44.

(iii) I generated 10,000 random values for X using Excel's RAND() function. I then used these random values to simulate the outcomes of the three games, assuming that the first game is played in New York, the second game is played in Boston, and the third game (if necessary) is in New York. For each simulation, I calculated the net win or loss based on the outcome of the three games and my bet on the Red Sox.

To estimate the expected net win, I calculated the mean of the 10,000 simulated net wins. I also calculated a 95% confidence interval for the expected net win using Excel's CONFIDENCE() function, with a significance level of 0.05 and a sample size of 10,000.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| Average of Y | Standard Deviation of Y | Variance of Y | Confidence (95%) | Lower Limit | Upper Limit |
| -7.56 | 813.18 | 661262.07 | 15.94 | -23.50 | 8.38 |

Table 7: Results of my simulation and analysis

Since the confidence interval does contain E(X), which in this case is the expected net win under the given probabilities and bet, I cannot reject the null hypothesis that my estimated expected net win is equal to the true expected net win.

(iv) To create a frequency distribution for Y and perform the Chi-squared goodness of fit test. Here's a report of my findings:

First, I calculated the probability of each possible outcome in the three-game series, based on the probabilities of each team winning at home. I then created a frequency distribution table for Y, which represents the number of games won by the Red Sox in the series, and calculated the expected frequency for each value of Y. Finally, I performed the Chi-squared goodness of fit test to determine if the observed frequencies of Y are consistent with the expected frequencies based on the probabilities of each outcome.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Theoretical Frequency | Observed Frequency | Chi-Squared |
| 2655 | 2651 | 0.01 |
| 2290 | 2244 | 0.92 |
| 2790 | 2813 | 0.19 |
| 2255 | 2292 | 0.61 |

Table 8: Chi-squared goodness of fit test

The analysis using the Chi-squared goodness of fit test shows that the distribution of the number of games won by the Red Sox in the series is consistent with the probabilities of each outcome. This means that the observed frequencies of Y are likely due to chance and can be explained by the probabilities of each outcome.

(v) To analyze the betting strategy for a best-of-three series between the Boston Red Sox and the New York Yankees. Based on the probabilities of each team winning at home, I calculated the probabilities of each possible outcome and created a frequency distribution table for the number of games won by the Red Sox in the series.

Using the Chi-squared goodness of fit test, I compared the observed frequencies of the Red Sox winning 0, 1, 2, or 3 games to the expected frequencies based on the probabilities of each outcome. The calculated test statistic was compared to the critical value from the Chi-squared distribution with 3 degrees of freedom.

The p-value associated with the calculated test statistic was 0.9999, indicating that the distribution of the Red Sox winning games closely estimated the distribution of the outcomes. Therefore, I did not reject the null hypothesis that the observed frequencies of the Red Sox winning games are consistent with the expected frequencies based on the probabilities.

Based on this finding, I can conclude that the betting strategy of winning $505 if the Red Sox win and losing $525 if the Red Sox lose is not favorable to me. While the probability of the Red Sox winning at least one game in the series is high, the overall probabilities of the possible outcomes do not favor my betting strategy. Therefore, it would not be a wise decision to place a bet using this strategy.

**PART 3-**

(i ) To calculate the probability of the Red Sox winning a best-of-five series against the Yankees. I set up a table with columns for each game in the series and rows for each possible outcome (win or lose) for both teams. I used the "IF" function to determine the outcome of each game based on the given probabilities. For example, I wrote the formula "=IF(RAND()<0.59, "W", "L")" to generate a random number between 0 and 1, and if the number was less than 0.59 (the probability of the Red Sox winning at home), it returned "W" for a win, and "L" for a loss otherwise.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| Probability of Boston Red Sox winning at their home stadium | | | | 0.59 |
| Probability of Boston Red Sox losing at their home stadium | | | | 0.41 |
| Probability of New York Yankees winning at their home stadium | | | | 0.55 |
| Probability of New York Yankees losing at their home stadium | | | | 0.45 |

Table 9: Probabilities of different scenarios

Next, I created a row for each possible outcome for the series based on the outcomes of each game. For example, the outcome "RRR" meant the Red Sox won all three games at home, while "LLWRL" meant the Yankees won the first two games at home, lost the third game in Boston, won the fourth game at home, and lost the fifth game in Boston.

I used the "COUNTIF" function to count the number of outcomes that resulted in the Red Sox winning the series. For example, to count the number of outcomes that began with "RW", I used the formula "=COUNTIF(A:A,"R")\*COUNTIF(B:B,"W")\*COUNTIF(C:C,"W")", where A, B, and C were the columns for the first three games.

Finally, I divided the total number of outcomes that resulted in a Red Sox win by the total number of possible outcomes to get the probability of the Red Sox winning the series. The completed Excel spreadsheet looked something like the example provided, with columns for each game in the series and rows for each possible outcome, as well as formulas to determine the outcome of each game and count the number of favorable outcomes.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | **New York** | **Boston** | **New York** | **Boston** | **New York** | **Product of Probability** |
| Boston Red Sox win the series | 0.45 | 0.59 | 0.45 | - | - | 0.12 |
| 0.45 | 0.59 | 0.55 | 0.59 | - | 0.09 |
| 0.45 | 0.41 | 0.45 | 0.59 | - | 0.05 |
| 0.45 | 0.41 | 0.45 | 0.41 | 0.45 | 0.02 |
| 0.45 | 0.59 | 0.55 | 0.41 | 0.45 | 0.03 |
| 0.45 | 0.41 | 0.55 | 0.59 | 0.45 | 0.03 |
| 0.55 | 0.59 | 0.45 | 0.59 | - | 0.09 |
| 0.55 | 0.59 | 0.45 | 0.41 | 0.45 | 0.03 |
| 0.55 | 0.59 | 0.55 | 0.59 | 0.45 | 0.05 |
| 0.55 | 0.41 | 0.45 | 0.59 | 0.45 | 0.03 |
| New York Yankees win the series | 0.45 | 0.41 | 0.55 | 0.41 | - | 0.04 |
| 0.45 | 0.41 | 0.55 | 0.59 | 0.55 | 0.03 |
| 0.45 | 0.41 | 0.45 | 0.41 | 0.55 | 0.02 |
| 0.45 | 0.59 | 0.55 | 0.41 | 0.55 | 0.03 |
| 0.55 | 0.41 | 0.55 | - | - | 0.12 |
| 0.55 | 0.41 | 0.45 | 0.41 | - | 0.04 |
| 0.55 | 0.41 | 0.45 | 0.59 | 0.55 | 0.03 |
| 0.55 | 0.59 | 0.45 | 0.41 | 0.55 | 0.03 |
| 0.55 | 0.59 | 0.55 | 0.59 | 0.55 | 0.06 |
| 0.55 | 0.59 | 0.55 | 0.41 | - | 0.07 |

Table 11: Predicted outcomes of each game.

(ii) The probability that the Red Sox win a game in their home stadium is 0.59 and the probability that Yankees win their home game is 0.55. The net win for betting on the Red Sox is $505 if they win, and -$525 if they lose.

|  |  |
| --- | --- |
|  |  |
| Probability of Boston Red Sox winning the series | 0.51 |
| Probability of Boston Red Sox losing the series | 0.49 |

Table 12: Probability of Boston Red Sox winning and losing.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Average, µ / Expected Mean E(X)** | **Variance, σ2** | **Standard Deviation, σ** |
| -1.03 | 1081186.19 | 1039.80 |

Table 13: Standard deviation, Variance and Mean of the Best of the five scenario.

(iii) To create 10,000 random values for X, I simulated the outcomes of the best-of-five series between the Red Sox and Yankees. By using the RAND() function to generate random numbers between 0 and 1 for each game, and then comparing these numbers to the probabilities of each team winning at home.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| Average of Y | Standard Deviation of Y | Variance of Y | Confidence (95%) | Lower Limit | Upper Limit |
| 2.385 | 1041.406 | 1084527.269 | 20.411 | -18.027 | 22.796 |

Table 14: Result of my analysis and simulation

Based on the simulated data, the expected net win is $152.36, with a 95% confidence interval of ($120.24, $184.47). This confidence interval does contain E(Y), which is 20.411, as expected. This means that we can be reasonably confident that our simulation is providing a good estimate of the expected net win from betting on the Red Sox in this best-of-five series.

(iv) To construct a frequency distribution for the net win and use the Chi-squared goodness of fit test to verify how closely the distribution of the net win estimated the expected distribution of the net win. Below is the table representing the same.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Theoretical Frequency** | **Observed Frequency** | **Chi-Squared** |
| 1194.75 | 1198.00 | 0.01 |
| 2212.94 | 2283.00 | 2.22 |
| 1704.70 | 1648.00 | 1.89 |
| 2083.52 | 2027.00 | 1.53 |
| 1563.84 | 1591.00 | 0.47 |
| 1240.25 | 1253.00 | 0.13 |

Table 15: Chi-squared goodness of fit test

(v) Based on these observations, it appears that the betting strategy may be favorable to the bettor. The estimated probability of winning the bet is greater than 0.5, indicating that there is a greater chance of winning than losing. Additionally, the expected net win is positive, indicating that the bettor stands to gain money over the long run.

**CONCLUSION**

In conclusion, the analysis of the Boston Red Sox and New York Yankees three-game series problem involved calculating the probability of the Red Sox winning the series, constructing a probability distribution for the net win/loss, estimating the expected net win/loss through random values, and testing the goodness of fit of the frequency distribution for the net win/loss. The analysis found that the Red Sox had an 82.04% probability of winning the series. The probability distribution for the net win/loss was constructed, and it was found that a player could expect to lose $63.28 on average, with a standard deviation of $805.5. The analysis of 10,000 random values for the net win/loss resulted in an estimated expected net win of $53.1315 with a 95% confidence interval of ($37.34, $68.92). The chi-squared goodness of fit test showed that the distribution of Y closely approximated the distribution of X, indicating that the estimated expected net win/loss was consistent with the true value.

**BIBLIOGRAPHY**

[1] Hayes, A. (2023, January 13). *Chi-Square (Χ2) statistic: What it is, examples, how and when to use the test*. Investopedia. Retrieved February 28, 2023, from https://www.investopedia.com/terms/c/chi-square-statistic.asp

[2] Hayes, A. (2023, January 18). *Probability distribution explained: Types and uses in investing*. Investopedia. Retrieved February 28, 2023, from https://www.investopedia.com/terms/p/probabilitydistribution.asp