MODULE ONE PROJECT:

ANALYSIS OF A BETTING STRATEGY IN SPORTS

By : **HARSHIT GAUR**
MASTER OF PROFESSIONAL STUDIES IN ANALYTICS
ALY 6050 : INTRODUCTION TO ENTERPRISE ANALYTICS
JANUARY 16, 2022

To : **PROF. RICHARD HE**

ABSTRACT

Data analytics is a discipline which is focused on extracting meaningful insights from data. It comprises of processes, techniques, and tools of data analysis and management, including the collection, organization, and storage of data. The primary aim of data analytics is to apply statistical analysis on data in order to find trends and solve problems.

Probability is a science of the likeliness of events to happen. The probability of an event is a number ranging between 0 and 1, where, if we speak roughly, the value 0 indicates impossibility of the event to happen while the value 1 indicates the certainty of the event to happen. The higher the probability, the more likely it is that the event will occur.

INTRODUCTION

In this project, I will be applying Probability Theory to analyse the betting strategies in sports. We will be analysing the betting strategies in a series of Baseball games between Boston's Red Sox and New York's Yankees. This will help us to understand the probability theory and its applicability in order to find the probability of likelihood of an event.

PROBLEM STATEMENT

Suppose that Boston Red Sox and New York Yankees (two American League baseball team) are scheduled to play series of games. There are 4 different types of series to be played between them. The probability that the Red Sox win a game at their home ground is 0.59 and the probability that the Yankees win a game at their home ground is 0.56. The winner of the series will be the first team who wins :

a. Two games out of three games in a **Best-of-Three** series of games. The first game is played in New York.
b. Two games out of three games in a **Best-of-Three** series of games. The first game is played in Boston.
c. Three games out of five games in a **Best-of-Five** series of games.
d. Four games out of seven games in a **Best-of-Seven** series of games.

Next, suppose that I have placed a bet on each game played where if **Red Sox wins a game**, I earn $**1000** and if **Red Sox loses a game**, I lose $**1050.** The outcomes of the games are independent of each other in every case.

All the different cases of betting strategy will be discussed, and analysed, The probability distributions of net wins (X) will also be listed.

Next, I will generate 10,000 random values from the value pool of net wins (X) calculated in the above step. These values will be used to calculate the 95% Confidence Interval (CI) and the expected net win E(X).

Later, the frequency distribution of Observed frequencies (randomly generated) and Theoretical frequencies (estimated probabilities) will be generated and compared against each other using Chi-Squared Goodness of Fit test.

Part 1: Best of Three Series (*First Game in New York*)

The first game is played in New York at the home ground of New York Yankees. The second game is played in Boston at the home ground of Boston Red Sox. The third game (if becomes necessary) is played in New York.

I. Calculate the probability that Boston Red Sox will win the series.

New York	Boston	New York
WIN	WIN	Not Played
WIN	LOSE	WIN
LOSE	WIN	WIN
LOSE	LOSE	Not Played
LOSE	WIN	LOSE
WIN	LOSE	LOSE

Figure 1.1: Sample space of the Series.

New York	Boston	New York	Product of Probability
0.44	0.59	-	0.2596
0.44	0.41	0.44	0.079376
0.56	0.59	0.44	0.145376
0.56	0.41	-	0.2296
0.56	0.59	0.56	0.185024
0.44	0.41	0.56	0.101024

Figure 1.2: Product of Probabilities of games.

Probability of Boston Red Sox winning the series	**0.484352**
Probability of Boston Red Sox losing the series	0.515648
Sum of Probabilities	1

Figure 1.3: Probabilities of Boston Red Sox winning and losing the series.

1. There are 6 cases of series of games in which Boston's Red Sox wins 3 series and the rest are won by New York's Yankees. Each team has to win 2 games out of 3 games in this **Best-of-three** series.

2. The series of games where Boston wins the series are :
    1. Win, Win
    2. Win, Lose, Win
    3. Lose, Win, Win

3. The probability of Boston's Red Sox winning the series of **Best-of-Three** is **0.484352**.

4. The probability of Boston's Red Sox losing the series of **Best-of-Three** is **0.515648**.

II. Construct a probability distribution for your net win (X) in the series. Calculate your expected net win (the mean of X) and the standard deviation of X.

Winnings / Losings on each game in series			Net Wins
1000	1000	0	2000
1000	-1050	1000	950
-1050	1000	1000	950
-1050	-1050	0	-2100
-1050	1000	-1050	-1100
1000	-1050	-1050	-1100

Figure 1.4: Winnings/Losing on each game in series.

WINNINGS ($) X	P(X)	CUMMULATIVE P(X)	X^2
2000	0.2596	0.2596	4000000
950	0.224752	0.484352	902500
-1100	0.286048	0.7704	1210000
-2100	0.2296	1	4410000

Figure 1.5: Frequency Distribution of the series.

FINDINGS	
Average, μ / Expected Mean E(X) =	-64.10
Variance, σ2 =	2595784.16
Standard Deviation, σ =	1611.14

Figure 1.6: Average, Variance, Standard Deviation

1. The person earns $1000 when Boston wins a game and loses $1050 when Boston loses a game in the series. A table of Net Wins (X) is generated with the earnings and losings after a team has won the series.

2. The probability distribution for Net Wins (X) in the series, generated in the above step, is generated using functions provided by Excel.

3. Average (μ) or Expected Mean E(X), Variance, and Standard Deviation are calculated using the probability distribution.

4. Expected Mean value E(X) is equal to **-64.10** and Standard Deviation is **1611.14.**

III) Create 10,000 random variables for X and use them to estimate expected net win by using 95% confidence interval. Check if this confidence interval contain E(X)?

Random Values	
Rand()	**Y**
0.519253217	-1100
0.590151035	-1100
0.935167103	-2100
0.856449156	-2100
0.847577528	-2100
0.976843459	-2100
0.384439747	950
0.930449086	-2100
0.501915092	-1100
0.453057307	950
0.64537697	-1100

Figure 1.7: Random values generated using Rand() and VLOOKUP() function.

VLOOKUP	
CUMMULATIVE P(X)	**WINNINGS ($) X**
0	2000
0.2596	950
0.484352	-1100
0.7704	-2100

Figure 1.8: VLOOKUP table of Net earnings (winnings/ losings).

FINDINGS	
Average of Y =	-53.625
Variance of Y =	2572884.898
Standard Deviation of Y =	1604.021477
Confidence (95%) =	31.43824325
Lower Limit =	-85.06324325
Upper Limit =	-22.18675675

Figure 1.9: Findings of the randomly generated series.

1. 10,000 random variables for X are generated using RAND() function and mapped these random decimal values to original values using **VLOOKUP()** function. This series is given the name of Y.

2. Statistics of this series are calculated using functions provided in Excel tool and Confidence Interval is calculated using Confidence Value (provided via CONFIDENCE() function).

3. The Mean (Average of Y) calculated is **-53.625**, Standard Deviation (SD) is **1604.02** and Confidence value (Margin of Error) at 95% is **31.438.**

4. The **Lower Limit of Confidence Interval** is **-85.063**, and **Upper Limit of Confidence Interval** is **-22.186.**

IV) Construct Frequency Distribution for Y (randomly generated 10,000 values). Use Chi-Squared Goodness of Fit test to verify how closely the distribution of Y has estimated the distribution of X.

Theoretical Frequency	Observed Frequency	Chi-Squared
2596	2576	0.154083205
2247.52	2192	1.371498541
2860.48	2853	0.019559794
2296	2379	3.00043554
	Degree of Freedom	Chi-Squared Metric
	3	4.545577
	P-value	P-value using Chi-Square Test Function
	0.208	0.208

Figure 1.10: Chi-Squared Goodness of Fit test applied to theoretical and observed frequencies.

1. Theoretical and Observed Frequencies are calculated for the Net Wins (X) and Chi-Squared values are calculated for each of the Net Win values.

2. These Chi-Squared Metrics are summed up to calculate the final chi-square value.

3. Using this final Chi-Squared Metric and Degree of Freedom, P-value is calculated using the function **CHISQ.TEST()** function.

4. The NULL Hypothesis,
   H_0 : Theoretical Frequency distribution = Observed Frequency distribution

5. The ALTERNATE Hypothesis,
   H_1 : Theoretical Frequency distribution ≠ Observed Frequency distribution

6. Alpha value = 0.05 (95% confidence). To verify the P-value, the series of theoretical and observed frequencies are used in the function **CHISQ.DIST().**

7. The P-value of the series comes out to be **0.208**. This P-value changes every time whenever random values are changed at each iteration. But, for this iteration, it is 0.208.

V. Describe whether betting strategy is favourable to us or not using the observations above in the case.

- According to the observations, **the betting strategy is NOT good for me.**
- With 95% confidence, I can say that the mean value of my total earnings would lie between the values **-85.063** and **-22.186.**
- These ***values are negative and signifies that I would lose money*** if I place bet on this series.

Part 2: Best of Three Series (*First Game in Boston*)

The first game is played in Boston at the home ground of Boston's Red Sox. The second game is played in New York at the home ground of New York Yankees. The third game (if becomes necessary) is played in Boston.

I. Calculate the probability that Boston Red Sox will win the series.

Boston	New York	Boston
WIN	WIN	Not Played
WIN	LOSE	WIN
LOSE	WIN	WIN
LOSE	LOSE	Not Played
LOSE	WIN	LOSE
WIN	LOSE	LOSE

Figure 2.1: Sample space of the Series.

Boston	New York	Boston	Product of Probability
0.59	0.44	-	0.2596
0.59	0.56	0.59	0.194936
0.41	0.44	0.59	0.106436
0.41	0.56	-	0.2296
0.41	0.44	0.41	0.073964
0.59	0.56	0.41	0.135464

Figure 2.2: Product of Probabilities of games.

Probability of Boston Red Sox winning the series	**0.560972**
Probability of Boston Red Sox losing the series	0.439028
Sum of Probabilities	1

Figure 2.3: Probabilities of Boston Red Sox winning and losing the series.

1. There are 6 cases of series of games in which Boston's Red Sox wins 3 series and the rest 3 are won by New York's Yankees. Each team has to win 2 games out of 3 games in this **Best-of-three** series.

2. The series of games where Boston wins the series are :
    1. Win, Win
    2. Win, Lose, Win
    3. Lose, Win, Win

3. The probability of Boston's Red Sox winning the series of **Best-of-Three** is **0.560972**.

4. The probability of Boston's Red Sox losing the series of **Best-of-Three** is **0.439028**.

II. Construct a probability distribution for your net win (X) in the series. Calculate your expected net win (the mean of X) and the standard deviation of X.

Winnings / Losings on each game in series			Net Wins
1000	1000	0	2000
1000	-1050	1000	950
-1050	1000	1000	950
-1050	-1050	0	-2100
-1050	1000	-1050	-1100
1000	-1050	-1050	-1100

Figure 2.4: Winnings/Losing on each game in series.

WINNINGS ($) X	P(X)	CUMMULATIVE P(X)	X^2
2000	0.2596	0.2596	4000000
950	0.301372	0.560972	902500
-1100	0.209428	0.7704	1210000
-2100	0.2296	1	4410000

Figure 2.5: Frequency Distribution of the series.

FINDINGS	
Average, μ / Expected Mean E(X) =	92.97
Variance, σ2 =	2567688.21
Standard Deviation, σ =	1602.40

Figure 2.6: Average, Variance, Standard Deviation

1. The person earns $1000 when Boston wins a game and loses $1050 when Boston loses a game in the series. A table of Net Wins (X) is generated with the earnings and losings after a team has won the series.

2. The probability distribution for Net Wins (X) in the series, generated in the above step, is generated using functions provided by Excel.

3. Average (μ) or Expected Mean E(X), Variance, and Standard Deviation are calculated using the probability distribution.

4. Expected Mean value E(X) is equal to **92.97** and Standard Deviation is **1602.40.**

III) Create 10,000 random variables for X and use them to estimate expected net win by using 95% confidence interval. Check if this confidence interval contain E(X)?

Random Values	
Rand()	**Y**
0.410510999	950
0.619088276	-1100
0.840099015	-2100
0.598694451	-1100
0.584963161	-1100
0.387448392	950
0.096402503	2000
0.338421055	950
0.384678651	950
0.11326898	2000
0.866286327	-2100
0.751050984	-1100
0.907624026	-2100

Figure 2.7: Random values generated using Rand() and VLOOKUP() function. This series is different on each iteration.

VLOOKUP	
CUMMULATIVE P(X)	**WINNINGS ($) X**
0	2000
0.2596	950
0.560972	-1100
0.7704	-2100

Figure 2.8: VLOOKUP table of Net earnings (winnings/ losings).

FINDINGS	
Average of Y =	94.58
Variance of Y =	2582274.351
Standard Deviation of Y =	1606.945659
Confidence Value (95%) =	31.49555617
Lower Limit =	63.08444383
Upper Limit =	126.0755562

Figure 2.9: Findings of the randomly generated series.

1. 10,000 random variables for X are generated using RAND() function and mapped these random decimal values to original values using **VLOOKUP()** function. This series is given the name of Y.

2. Statistics of this series are calculated using functions provided in Excel tool and Confidence Interval is calculated using Confidence Value (provided via CONFIDENCE() function).

3. The Mean (Average of Y) calculated is **94.58**, Standard Deviation (SD) is **1606.94** and Confidence value (Margin of Error) at 95% is **31.495.**

4. The **Lower Limit of Confidence Interval** is **63.084**, and **Upper Limit of Confidence Interval** is **126.075.**

IV) Construct Frequency Distribution for Y (randomly generated 10,000 values). Use Chi-Squared Goodness of Fit test to verify how closely the distribution of Y has estimated the distribution of X.

Theoretical Frequency	Observed Frequency	Chi-Squared
2596	2593	0.003466872
3013.72	3020	0.013086285
2094.28	2068	0.329773669
2296	2319	0.230400697
	Degree of Freedom	Chi-Squared Metric
	3	0.576728
	P-value	P-value using Chi-Square Test Function
	0.902	0.902

Figure 2.10: Chi-Squared Goodness of Fit test applied to theoretical and observed frequencies.

1. Theoretical and Observed Frequencies are calculated for the Net Wins (X) and Chi-Squared values are calculated for each of the Net Win values.

2. These Chi-Squared Metrics are summed up to calculate the final chi-square value.

3. Using this final Chi-Squared Metric and Degree of Freedom, P-value is calculated using the function **CHISQ.TEST()** function.

4. The NULL Hypothesis,
   H_0 : Theoretical Frequency distribution = Observed Frequency distribution

5. The ALTERNATE Hypothesis,
   H_1 : Theoretical Frequency distribution ≠ Observed Frequency distribution

6. Alpha value = 0.05 (95% confidence). To verify the P-value, the series of theoretical and observed frequencies are used in the function **CHISQ.DIST().**

7. The P-value of the series comes out to be **0.902**. This P-value changes every time whenever random values are changed at each iteration. But, for this iteration, it is 0.902.

V. Describe whether betting strategy is favourable to us or not using the observations above in the case.

- According to the observations, **the betting strategy is good for me.**
- With 95% confidence, I can say that the mean value of my total earnings would lie between the values **63.084** and **126.075.**
- These ***values are positive and signifies that I would earn money*** if I place bet on this series. Although, the earnings are not very high, but I'll be in profit.

Part 3: Best of Five Series (*First Game in New York*)

The first game is played in New York at the home ground of New York Yankees. The second game is played in Boston at the home ground of Boston Red Sox. The third game (if becomes necessary) is played in New York.

I. Calculate the probability that Boston Red Sox will win the series.

New York	Boston	New York	Boston	New York
WIN	WIN	WIN	Not Played	Not Played
WIN	WIN	LOSE	WIN	Not Played
WIN	LOSE	WIN	WIN	Not Played
WIN	LOSE	WIN	LOSE	WIN
WIN	WIN	LOSE	LOSE	WIN
WIN	LOSE	LOSE	WIN	WIN
LOSE	WIN	WIN	WIN	Not Played
LOSE	WIN	WIN	LOSE	WIN
LOSE	WIN	LOSE	WIN	WIN
LOSE	LOSE	WIN	WIN	WIN
WIN	LOSE	LOSE	LOSE	Not Played
WIN	LOSE	LOSE	WIN	LOSE
WIN	LOSE	WIN	LOSE	LOSE
WIN	WIN	LOSE	LOSE	LOSE
LOSE	LOSE	LOSE	Not Played	Not Played
LOSE	LOSE	WIN	LOSE	Not Played
LOSE	LOSE	WIN	WIN	LOSE
LOSE	WIN	WIN	LOSE	LOSE
LOSE	WIN	LOSE	WIN	LOSE
LOSE	WIN	LOSE	LOSE	Not Played

Figure 3.1: Sample space of the Series.

New York	Boston	New York	Boston	New York	Product of Probability
0.44	0.59	0.44	-	-	0.114224
0.44	0.59	0.56	0.59	-	0.08577184
0.44	0.41	0.44	0.59	-	0.04683184
0.44	0.41	0.44	0.41	0.44	0.01431943
0.44	0.59	0.56	0.41	0.44	0.02622583
0.44	0.41	0.56	0.59	0.44	0.02622583
0.56	0.59	0.44	0.59	-	0.08577184
0.56	0.59	0.44	0.41	0.44	0.02622583
0.56	0.59	0.56	0.59	0.44	0.04803223
0.56	0.41	0.44	0.59	0.44	0.02622583
0.44	0.41	0.56	0.41	-	0.04141984
0.44	0.41	0.56	0.59	0.56	0.03337833
0.44	0.41	0.44	0.41	0.56	0.01822473
0.44	0.59	0.56	0.41	0.56	0.03337833
0.56	0.41	0.56	-	-	0.128576
0.56	0.41	0.44	0.41	-	0.04141984
0.56	0.41	0.44	0.59	0.56	0.03337833
0.56	0.59	0.44	0.41	0.56	0.03337833
0.56	0.59	0.56	0.59	0.56	0.06113193
0.56	0.59	0.56	0.41	-	0.07585984

Figure 3.2: Product of Probabilities of games.

Probability of Boston Red Sox winning the series	**0.499854502**
Probability of Boston Red Sox losing the series	0.500145498
Sum of Probabilities	1

Figure 3.3: Probabilities of Boston Red Sox winning and losing the series.

1. There are 20 cases of series of games in which Boston's Red Sox wins 10 series and the rest 10 are won by New York's Yankees. Each team has to win 3 games out of 3 games in this **Best-of-Five** series.

2. The probability of Boston's Red Sox winning the series of **Best-of-Five** is **0.4998545**.

3. The probability of Boston's Red Sox losing the series of **Best-of-Five** is **0.5001454**.

II. Construct a probability distribution for your net win (X) in the series. Calculate your expected net win (the mean of X) and the standard deviation of X.

Winnings / Losings on each game in series					Net Wins
1000	1000	1000	0	0	3000
1000	1000	-1050	1000	0	1950
1000	-1050	1000	1000	0	1950
1000	-1050	1000	-1050	1000	900
1000	1000	-1050	-1050	1000	900
1000	-1050	-1050	1000	1000	900
-1050	1000	1000	1000	0	1950
-1050	1000	1000	-1050	1000	900
-1050	1000	-1050	1000	1000	900
-1050	-1050	1000	1000	1000	900
1000	-1050	-1050	-1050	0	-2150
1000	-1050	-1050	1000	-1050	-1150
1000	-1050	1000	-1050	-1050	-1150
1000	1000	-1050	-1050	-1050	-1150
-1050	-1050	-1050	0	0	-3150
-1050	-1050	1000	-1050	0	-2150
-1050	-1050	1000	1000	-1050	-1150
-1050	1000	1000	-1050	-1050	-1150
-1050	1000	-1050	1000	-1050	-1150
-1050	1000	-1050	-1050	0	-2150

Figure 3.4: Winnings/Losing on each game in series.

WINNINGS ($) X	P(X)	CUMMULATIVE P(X)	X^2
3000	0.114224	0.114224	9000000
1950	0.21837552	0.33259952	3802500
900	0.167254982	0.499854502	810000
-1150	0.212869978	0.71272448	1322500
-2150	0.15869952	0.871424	4622500
-3150	0.128576	1	9922500

Figure 3.5: Frequency Distribution of the series.

FINDINGS	
Average, μ / Expected Mean E(X) =	-71.99
Variance, σ2 =	4279588.03
Standard Deviation, σ =	2068.72

Figure 3.6: Average, Variance, Standard Deviation

1. The person earns $1000 when Boston wins a game and loses $1050 when Boston loses a game in the series. A table of Net Wins (X) is generated with the earnings and losings after a team has won the series.
2. The probability distribution for Net Wins (X) in the series, generated in the above step, is generated using functions provided by Excel.
3. Average (μ) or Expected Mean E(X), Variance, and Standard Deviation are calculated using the probability distribution.

4. Expected Mean value E(X) is equal to **-71.99** and Standard Deviation is **2068.72.**

III) Create 10,000 random variables for X and use them to estimate expected net win by using 95% confidence interval. Check if this confidence interval contain E(X)?

Random Values	
Rand()	**Y**
0.767196215	-2150
0.582613406	-1150
0.14386938	1950
0.609558381	-1150
0.669773116	-1150
0.262478596	1950
0.523430334	-1150
0.102547311	3000
0.529860502	-1150
0.150682957	1950
0.259528358	1950
0.514570506	-1150

Figure 3.7: Random values generated using Rand() and VLOOKUP() function. This series is different on each iteration.

VLOOKUP	
CUMULATIVE P(X)	**WINNINGS ($) X**
0	3000
0.114224	1950
0.33259952	900
0.499854502	-1150
0.71272448	-2150
0.871424	-3150

Figure 3.8: VLOOKUP table of Net earnings (winnings/ losings).

FINDINGS	
Average of Y =	-71.44
Variance of Y =	4319910.817
Standard Deviation of Y =	2078.439515
Confidence (95%) =	40.73666593
Lower Limit =	-112.1766659
Upper Limit =	-30.70333407

Figure 3.9: Findings of the randomly generated series.

1. 10,000 random variables for X are generated using RAND() function and mapped these random decimal values to original values using **VLOOKUP()** function. This series is given the name of Y.

2. Statistics of this series are calculated using functions provided in Excel tool and Confidence Interval is calculated using Confidence Value (provided via CONFIDENCE() function).

3. The Mean (Average of Y) calculated is **-71.44**, Standard Deviation (SD) is **2078.43** and Confidence value (Margin of Error) at 95% is **40.736.**

4. The **Lower Limit of Confidence Interval** is **-112.176**, and **Upper Limit of Confidence Interval** is **-30.703.**

IV) Construct Frequency Distribution for Y (randomly generated 10,000 values). Use Chi-Squared Goodness of Fit test to verify how closely the distribution of Y has estimated the distribution of X.

Theoretical Frequency	Observed Frequency	Chi-Squared
1142.24	1147	0.019836112
2183.7552	2162	0.216731586
1672.549824	1722	1.462031128
2128.699776	2097	0.472060837
1586.9952	1610	0.333473487
1285.76	1262	0.439069189
	Degree of Freedom	Chi-Squared Metric
	5	2.943202
	P-value	P-value using Chi-Square Test Function
	0.709	0.709

Figure 3.10: Chi-Squared Goodness of Fit test applied to theoretical and observed frequencies.

1. Theoretical and Observed Frequencies are calculated for the Net Wins (X) and Chi-Squared values are calculated for each of the Net Win values.

2. These Chi-Squared Metrics are summed up to calculate the final chi-square value.

3. Using this final Chi-Squared Metric and Degree of Freedom, P-value is calculated using the function **CHISQ.TEST()** function.

4. The NULL Hypothesis,
   H_0 : Theoretical Frequency distribution = Observed Frequency distribution

5. The ALTERNATE Hypothesis,
   H_1 : Theoretical Frequency distribution ≠ Observed Frequency distribution

6. Alpha value = 0.05 (95% confidence). To verify the P-value, the series of theoretical and observed frequencies are used in the function **CHISQ.DIST().**

7. The P-value of the series comes out to be **0.709**. This P-value changes every time whenever random values are changed at each iteration. But, for this iteration, it is 0.709.

V. Describe whether betting strategy is favourable to us or not using the observations above in the case.

- According to the observations, **the betting strategy is NOT good for me.**
- With 95% confidence, I can say that the mean value of my total earnings would lie between the values **-112.176** and **-30.703**
- These ***values are negative and signifies that I would lose money*** if I place bet on this series.

Part 4: Best of Seven Series (*First Game in Boston*)

The first game is played in Boston at the home ground of Boston Red Sox. The second game is played in New York at the home ground of New York Yankees. The third game (if becomes necessary) is played in Boston Red Sox.

I. Calculate the probability that Boston Red Sox will win the series.

Boston	New York	Boston	New York	Boston	New York	Boston
WIN	WIN	WIN	WIN	Not Played	Not Played	Not Played
LOSE	WIN	WIN	WIN	WIN	Not Played	Not Played
WIN	LOSE	WIN	WIN	WIN	Not Played	Not Played
WIN	WIN	LOSE	WIN	WIN	Not Played	Not Played
WIN	WIN	WIN	LOSE	WIN	Not Played	Not Played
LOSE	LOSE	WIN	WIN	WIN	WIN	Not Played
LOSE	WIN	LOSE	WIN	WIN	WIN	Not Played
LOSE	WIN	WIN	LOSE	WIN	WIN	Not Played
LOSE	WIN	WIN	WIN	LOSE	WIN	Not Played
WIN	LOSE	LOSE	WIN	WIN	WIN	Not Played
WIN	LOSE	WIN	LOSE	WIN	WIN	Not Played
WIN	LOSE	WIN	WIN	LOSE	WIN	Not Played
WIN	WIN	LOSE	LOSE	WIN	WIN	Not Played
WIN	WIN	LOSE	WIN	LOSE	WIN	Not Played
WIN	WIN	WIN	LOSE	LOSE	WIN	Not Played
LOSE	LOSE	LOSE	WIN	WIN	WIN	WIN
LOSE	LOSE	WIN	LOSE	WIN	WIN	WIN
LOSE	LOSE	WIN	WIN	LOSE	WIN	WIN
LOSE	LOSE	WIN	WIN	WIN	LOSE	WIN
LOSE	WIN	LOSE	LOSE	WIN	WIN	WIN
LOSE	WIN	LOSE	WIN	LOSE	WIN	WIN
LOSE	WIN	LOSE	WIN	WIN	LOSE	WIN
LOSE	WIN	WIN	LOSE	LOSE	WIN	WIN
LOSE	WIN	WIN	LOSE	WIN	LOSE	WIN
LOSE	WIN	WIN	WIN	LOSE	LOSE	WIN
WIN	LOSE	LOSE	LOSE	WIN	WIN	WIN
WIN	LOSE	LOSE	WIN	LOSE	WIN	WIN
WIN	LOSE	LOSE	WIN	WIN	LOSE	WIN
WIN	LOSE	WIN	LOSE	LOSE	WIN	WIN
WIN	LOSE	WIN	LOSE	WIN	LOSE	WIN
WIN	LOSE	WIN	WIN	LOSE	LOSE	WIN
WIN	WIN	LOSE	LOSE	LOSE	WIN	WIN
WIN	WIN	LOSE	LOSE	WIN	LOSE	WIN
WIN	WIN	LOSE	WIN	LOSE	LOSE	WIN
WIN	WIN	WIN	LOSE	LOSE	LOSE	WIN
LOSE	LOSE	LOSE	LOSE	Not Played	Not Played	Not Played
WIN	LOSE	LOSE	LOSE	LOSE	Not Played	Not Played
LOSE	WIN	LOSE	LOSE	LOSE	Not Played	Not Played
LOSE	LOSE	WIN	LOSE	LOSE	Not Played	Not Played

Figure 4.1: Sample space of the Series.

Boston	New York	Boston	New York	Boston	New York	Boston	Product of Probability
0.59	0.44	0.59	0.44	-	-	-	0.06739216
0.41	0.44	0.59	0.44	0.59	-	-	0.027630786
0.59	0.56	0.59	0.44	0.59	-	-	0.050605386
0.59	0.44	0.41	0.44	0.59	-	-	0.027630786
0.59	0.44	0.59	0.56	0.59	-	-	0.050605386
0.41	0.56	0.59	0.44	0.59	0.44	-	0.01547324
0.41	0.44	0.41	0.44	0.59	0.44	-	0.008448464
0.41	0.44	0.59	0.56	0.59	0.44	-	0.01547324
0.41	0.44	0.59	0.44	0.41	0.44	-	0.008448464
0.59	0.56	0.41	0.44	0.59	0.44	-	0.01547324
0.59	0.56	0.59	0.56	0.59	0.44	-	0.028339016
0.59	0.56	0.59	0.44	0.41	0.44	-	0.01547324
0.59	0.44	0.41	0.56	0.59	0.44	-	0.01547324
0.59	0.44	0.41	0.44	0.41	0.44	-	0.008448464
0.59	0.44	0.59	0.56	0.41	0.44	-	0.01547324
0.41	0.56	0.41	0.44	0.59	0.44	0.59	0.006344028
0.41	0.56	0.59	0.56	0.59	0.44	0.59	0.011618997
0.41	0.56	0.59	0.44	0.41	0.44	0.59	0.006344028
0.41	0.56	0.59	0.44	0.59	0.56	0.59	0.011618997
0.41	0.44	0.41	0.56	0.59	0.44	0.59	0.006344028
0.41	0.44	0.41	0.44	0.41	0.44	0.59	0.00346387
0.41	0.44	0.41	0.44	0.59	0.56	0.59	0.006344028
0.41	0.44	0.59	0.56	0.41	0.44	0.59	0.006344028
0.41	0.44	0.59	0.56	0.59	0.56	0.59	0.011618997
0.41	0.44	0.59	0.44	0.41	0.56	0.59	0.006344028
0.59	0.56	0.41	0.56	0.59	0.44	0.59	0.011618997
0.59	0.56	0.41	0.44	0.41	0.44	0.59	0.006344028
0.59	0.56	0.41	0.44	0.59	0.56	0.59	0.011618997
0.59	0.56	0.59	0.56	0.41	0.44	0.59	0.011618997
0.59	0.56	0.59	0.56	0.59	0.56	0.59	0.021280025
0.59	0.56	0.59	0.44	0.41	0.56	0.59	0.011618997
0.59	0.44	0.41	0.56	0.41	0.44	0.59	0.006344028
0.59	0.44	0.41	0.56	0.59	0.56	0.59	0.011618997
0.59	0.44	0.41	0.44	0.41	0.56	0.59	0.006344028
0.59	0.44	0.59	0.56	0.41	0.56	0.59	0.011618997
0.41	0.56	0.41	0.56	-	-	-	0.05271616
0.59	0.56	0.41	0.56	0.41	-	-	0.031102534
0.41	0.44	0.41	0.56	0.41	-	-	0.016982134
0.41	0.56	0.59	0.56	0.41	-	-	0.031102534

Figure 4.2: Product of Probabilities of games.

Probability of Boston Red Sox winning the series	0.556799469
Probability of Boston Red Sox losing the series	0.443200531
Sum of Probabilities	1

Figure 3.3: Probabilities of Boston Red Sox winning and losing the series.

1. There are 70 cases of series of games in which Boston's Red Sox wins 35 series and the rest 35 are won by New York's Yankees. Each team has to win 3 games out of 4 games in this **Best-of-Seven** series.

2. The probability of Boston's Red Sox winning the series of **Best-of-Seven** is **0.556**.

3. The probability of Boston's Red Sox losing the series of **Best-of-Seven** is **0.443**.

II. Construct a probability distribution for your net win (X) in the series. Calculate your expected net win (the mean of X) and the standard deviation of X.

Winnings / Losings on each game in series							Net Wins
1000	1000	1000	1000	0	0	0	4000
-1050	1000	1000	1000	1000	0	0	2950
1000	-1050	1000	1000	1000	0	0	2950
1000	1000	-1050	1000	1000	0	0	2950
1000	1000	1000	-1050	1000	0	0	2950
-1050	-1050	1000	1000	1000	1000	0	1900
-1050	1000	-1050	1000	1000	1000	0	1900
-1050	1000	1000	-1050	1000	1000	0	1900
-1050	1000	1000	1000	-1050	1000	0	1900
1000	-1050	-1050	1000	1000	1000	0	1900
1000	-1050	1000	-1050	1000	1000	0	1900
1000	-1050	1000	1000	-1050	1000	0	1900
1000	1000	-1050	-1050	1000	1000	0	1900
1000	1000	-1050	1000	-1050	1000	0	1900
1000	1000	1000	-1050	-1050	1000	0	1900
-1050	-1050	-1050	1000	1000	1000	1000	850
-1050	-1050	1000	-1050	1000	1000	1000	850
-1050	-1050	1000	1000	-1050	1000	1000	850
-1050	-1050	1000	1000	1000	-1050	1000	850
-1050	1000	-1050	-1050	1000	1000	1000	850
-1050	1000	-1050	1000	-1050	1000	1000	850
-1050	1000	-1050	1000	1000	-1050	1000	850
-1050	1000	1000	-1050	-1050	1000	1000	850
-1050	1000	1000	-1050	1000	-1050	1000	850
-1050	1000	1000	1000	-1050	-1050	1000	850
1000	-1050	-1050	-1050	1000	1000	1000	850
1000	-1050	-1050	1000	-1050	1000	1000	850
1000	-1050	-1050	1000	1000	-1050	1000	850
1000	-1050	1000	-1050	-1050	1000	1000	850
1000	-1050	1000	-1050	1000	-1050	1000	850
1000	-1050	1000	1000	-1050	-1050	1000	850
1000	1000	-1050	-1050	-1050	1000	1000	850
1000	1000	-1050	-1050	1000	-1050	1000	850
1000	1000	-1050	1000	-1050	-1050	1000	850
1000	1000	1000	-1050	-1050	-1050	1000	850
-1050	-1050	-1050	-1050	0	0	0	-4200
1000	-1050	-1050	-1050	-1050	0	0	-3200
-1050	1000	-1050	-1050	-1050	0	0	-3200
-1050	-1050	1000	-1050	-1050	0	0	-3200
-1050	-1050	-1050	1000	-1050	0	0	-3200
1000	1000	-1050	-1050	-1050	-1050	0	-2200

Figure 4.4: Winnings/Losing on each game in series.

WINNINGS ($) X	P(X)	CUMULATIVE P(X)	X^2
4000	0.06739216	0.06739216	16000000
2950	0.156472342	0.223864502	8702500
1900	0.146523847	0.37038835	3610000
850	0.186411119	0.556799469	722500
-4200	0.05271616	0.609515629	17640000
-3200	0.096169338	0.705684966	10240000
-2200	0.164775103	0.87046007	4840000
-1200	0.12953993	1	1440000

Figure 4.5: Frequency Distribution of the series.

FINDINGS	
Average, μ / Expected Mean E(X) =	120.90
Variance, σ2 =	5987726.56
Standard Deviation, σ =	2446.98

Figure 4.6: Average, Variance, Standard Deviation

1. The person earns $1000 when Boston wins a game and loses $1050 when Boston loses a game in the series. A table of Net Wins (X) is generated with the earnings and losings after a team has won the series.

2. The probability distribution for Net Wins (X) in the series, generated in the above step, is generated using functions provided by Excel.

3. Average (μ) or Expected Mean E(X), Variance, and Standard Deviation are calculated using the probability distribution.

4. Expected Mean value E(X) is equal to **120.90** and Standard Deviation is **2446.98.**

III) Create 10,000 random variables for X and use them to estimate expected net win by using 95% confidence interval. Check if this confidence interval contain E(X)?

Random Values	
Rand()	**Y**
0.053504959	4000
0.0931403	2950
0.53053442	850
0.348678765	1900
0.741498318	-2200
0.85565937	-2200
0.57318207	-4200
0.822583779	-2200
0.088237593	2950
0.220154471	2950
0.940443904	-1200
0.458387393	850
0.621142904	-3200
0.134983009	2950

Figure 4.7: Random values generated using Rand() and VLOOKUP() function. This series is different on each iteration.

VLOOKUP	
CUMULATIVE P(X)	**WINNINGS ($) X**
0	4000
0.06739216	2950
0.223864502	1900
0.37038835	850
0.556799469	-4200
0.609515629	-3200
0.705684966	-2200
0.87046007	-1200

Figure 4.8: VLOOKUP table of Net earnings (winnings/ losings).

FINDINGS	
Average of Y =	100.645
Variance of Y =	6048105.645
Standard Deviation of Y =	2459.289663
Confidence (95%) =	48.20119166
Lower Limit =	52.44380834
Upper Limit =	148.8461917

Figure 4.9: Findings of the randomly generated series.

1. 10,000 random variables for X are generated using RAND() function and mapped these random decimal values to original values using **VLOOKUP()** function. This series is given the name of Y.

2. Statistics of this series are calculated using functions provided in Excel tool and Confidence Interval is calculated using Confidence Value (provided via CONFIDENCE() function).

3. The Mean (Average of Y) calculated is **-100.645**, Standard Deviation (SD) is **2459.28** and Confidence value (Margin of Error) at 95% is **48.201.**

4. The **Lower Limit of Confidence Interval** is **52.443**, and **Upper Limit of Confidence Interval** is **148.846.**

IV) Construct Frequency Distribution for Y (randomly generated 10,000 values). Use Chi-Squared Goodness of Fit test to verify how closely the distribution of Y has estimated the distribution of X.

Theoretical Frequency	Observed Frequency	Chi-Squared
673.9216	677	0.014061794
1564.723424	1583	0.213477491
1465.238474	1441	0.400961081
1864.111191	1790	2.946427564
527.1616	545	0.603626126
961.693376	981	0.387593114
1647.751034	1623	0.371787759
1295.399302	1360	3.221593658
	Degree of Freedom	Chi-Squared Metric
	7	8.159529
	P-value	P-value using Chi-Square Test Function
	0.319	0.319

Figure 4.10: Chi-Squared Goodness of Fit test applied to theoretical and observed frequencies.

1. Theoretical and Observed Frequencies are calculated for the Net Wins (X) and Chi-Squared values are calculated for each of the Net Win values.

2. These Chi-Squared Metrics are summed up to calculate the final chi-square value.

3. Using this final Chi-Squared Metric and Degree of Freedom, P-value is calculated using the function **CHISQ.TEST()** function.

4. The NULL Hypothesis,
   H_0 : Theoretical Frequency distribution = Observed Frequency distribution

5. The ALTERNATE Hypothesis,
   H_1 : Theoretical Frequency distribution ≠ Observed Frequency distribution

6. Alpha value = 0.05 (95% confidence). To verify the P-value, the series of theoretical and observed frequencies are used in the function **CHISQ.DIST().**

7. The P-value of the series comes out to be **0.319**. This P-value changes every time whenever random values are changed at each iteration. But, for this iteration, it is 0.319.

V. Describe whether betting strategy is favourable to us or not using the observations above in the case.

- According to the observations, **the betting strategy is good for me.**
- With 95% confidence, I can say that the mean value of my total earnings would lie between the values **52.443** and **148.846.**
- These ***values are positive and signifies that I would earn money*** if I place bet on this series.
- Even though the standard deviation is **2459.28**, the net wins seem to be highly volatile. But, the chances to earning money is also higher in this case. Therefore, I would go ahead with the betting strategy in this case.

CONCLUSION

We have applied Probability Theory to analyse the betting strategies in sports (in our case, Baseball). We have analysed the betting strategies in a series of Baseball games between Boston's Red Sox and New York's Yankees. The games are played in a series of **Three, Five, Seven** games differing in the venue of their first game (either Home ground or Away ground).

- According to the observations in a series of **Best-of-Three** games where the *first game is played in New York*, **the betting strategy is NOT good.** With 95% confidence, we can say that the mean value of my total earnings would lie between the values **-85.063** and **-22.186.** These ***values are negative and signifies that we would lose money,*** if we place bet on this series.

- According to the observations in a series of **Best-of-Three** games where the *first game is played in Boston*, **the betting strategy is actually good.** With 95% confidence, we can say that the mean value of my total earnings would lie between the values **63.084** and **126.075.** These ***values are positive and signifies that I would earn money*** if I place bet on this series. Although, the earnings are not very high, but I'll be in profit.

- According to the observations in a series of **Best-of-Five** games where the *first game is played in New York*, **the betting strategy is NOT good.** With 95% confidence, we can say that the mean value of my total earnings would lie between the values **-112.176** and **-30.703.** These ***values are negative and signifies that I would lose money*** if I place bet on this series.

- According to the observations in a series of **Best-of-Seven** games where the *first game is played in Boston*, **the betting strategy is actually good.** With 95% confidence, we can say that the mean value of my total earnings would lie between the values **52.443** and **148.846.** These ***values are positive and signifies that I would earn money*** if I place bet on this series. Even though the standard deviation is **2459.28**, the net wins seem to be highly volatile. But, the chances to earning money is also higher in this case. Therefore, I would go ahead with the betting strategy in this case.

- **We can also analyse that for a betting strategy to work in favour of the person if he/she/them is placing bet on Boston winning the series, the first game should be played in Boston at the home stadium of Boston Red Sox.**

BIBLIOGRAPHY

1. Microsoft. (2021). *Excel functions (alphabetical)*. https://support.microsoft.com/en-us/office/excel-functions-alphabetical-b3944572-255d-4efb-bb96-c6d90033e188

2. *Random number from fixed set of options*. (2022). ExcelJet | RandBetween. https://exceljet.net/formula/random-number-from-fixed-set-of-options

3. PerfectXL. (2021, August 24). *What is VLOOKUP // Excel glossary // PerfectXL Spreadsheet Validation*. https://www.perfectxl.com/excel-glossary/how-to-use-vlookup-excel/