INTRODUCTION

We are attempting to apply probability ideas to a betting scenario in this project. The findings of a research of the New York Yankees and Boston Red Sox (two American League baseball clubs) who are set to play a best-of-three series will be presented in this project. The series will be decided by the first side to win two of the three games, three of the five games & five of the seven games. The Red Sox have a 0.6 chance of winning a game at home, while the Yankees have a 0.57 chance of winning a game at home. Assume you put a wager on each game, winning $500 if the Red Sox win and losing $520 if the Red Sox lose.

Here are some of our team's winning conditions and probabilities, as well as the amount of money gained and lost per game. To determine the probability of our net winning amount, we must go over these and utilize the finest statistical methods, such as the chi-squared test, to obtain the best findings.

First task (Part One)

If the first game is played in New York, the second game is played in Boston, and the third game (if it becomes necessary) is in New York, then complete the following parts.

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

Here we calculated the value of Probability by writing down the condition in which Red Sox will win the series which was:

WW + WLW + LWW

Then we calculated values for each and added them together to get the following probability of Red Sox winning the series.

```
[1] "The probability that Red Sox will win the series = 0.4896"
```

Therefore, based on the above probability Red Sox has 48.96% of the winning the series if first game is played in New York, the second game is played in Boston, and the third game (if it becomes necessary) is in New York.

(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.

If the Red Sox win, we will collect $500, while if they lose, we will lose $520, according to our terms. So, for each match the probability of Red Sox winning and losing based on the home stadium of the team we get the net win and the standard deviation for the whole interaction

```
[1] "The expected net win = -50.4000000000001"
[1] "The standard deviation = 876.606776154508"
```

Here the net win is negative -50.40 which is expected as two of the three games will be played in New York home stadium where the probability of Yankees winning is way more than Red Sox.

(iii) Use R to create 10,000 random values for X. Let these random values be denoted by Y. Use these Y values to estimate your expected net win by using a 95% confidence interval. Does this confidence interval contain E(X)?

Here using the rnorm function on the net win and standard deviation which we obtained in the previous task, a range of random 10000 numbers were generated and then their mean and standard is calculated. Using the standard deviation and mean on the 95% confidence interval (which 1.96) the upper and lower limit was obtained.

```
[1] "The upper limit = -28.6627473828668"
[1] "The lower limit = -63.4450771453578"
```

Here we can notice that the upper and lower limit covers our net win of -50.40 value. Therefore, it means that the confidence interval contains our E(X) value.

(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.

A frequency distribution was constructed for Y which covers all the values of the outcome that can be obtained in the series. Which came out to be:

WW, WLW/LWW, LL, WLL/LWL

Where W = Win for Boston Red Sox & L = Win for New York Yankees.

Here taking out the profit/loss from the bets based on the game we get:

WW = 1000

WLW/LWW = 480

LL = -1040

WLL/LWL = -540

Then we get find the probability values for each outcome with their observed value which we can get from probability x 10000.

WW = 0.261 x 10000 = 2,610

WLW = 0.08505 x 10000 = 850.5

LL = 0.231 x 10000 = 2,310

WLL = 0.10395 x 10000 = 1039.5

LWL = 0.17545 x 10000 = 1754.5

LWW = 0.14355 x 10000 = 1435.5

Using the chi-square on the above observed values with their respective probability value we get

```
        Chi-squared test for given probabilities

data:  observed
X-squared = 4.4555, df = 3, p-value = 0.2163
```

Because frequency distributions of identical profit/loss values were summed (both their probability and their observed value) and displayed using the chi-square, the df is 3 and not 5.

(v) Use your observations of parts (ii) and (iii) above to describe whether your betting strategy is favorable to you.

The null hypothesis (Ho): states that the observed and experimental values acquired from the event are identical.

Alternative Hypothesis (H1): We found a discrepancy between observed and experimental values from the event.

Here we can notice that p-value is greater than 0.05(alpha value) therefore we don't have enough evidence to reject the null hypothesis meaning that observed and experimental values acquired from the event are identical.

Second task (Part Two)

Repeat part 1 above but assume that the first game is played in Boston, the second game is played in New York, and the third game (if it becomes necessary) is in Boston.

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

Here we calculated the value of Probability by writing down the condition in which Red Sox will win the series which was:

WW + WLW + LWW

Then we calculated values for each and added them together to get the following probability of Red Sox winning the series.

```
[1] "The probability that Red Sox will win the series = 0.55564"
```

Therefore, based on the above probability Red Sox has 55.564% of the winning the series if first game is played in Boston, the second game is played in New York, and the third game (if it becomes necessary) is in Boston.

(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.

If the Red Sox win, we will collect $500, while if they lose, we will lose $520, according to our terms. So, for each match the probability of Red Sox winning and losing based on the home stadium of the team we get the net win and the standard deviation for the whole interaction

```
[1] "The expected net win = 82.1999999999"
[1] "The standard deviation = 874.289357135268"
```

Here the net win is positive 82.19 which is expected as two of the three games will be played in Boston home stadium where the probability of Red Sox winning is way more than Yankees.

(iii) Use R to create 10,000 random values for X. Let these random values be denoted by Y. Use these Y values to estimate your expected net win by using a 95% confidence interval. Does this confidence interval contain E(X)?

Here using the rnorm function on the net win and standard deviation which we obtained in the previous task, a range of random 10000 numbers were generated and then their mean and standard is calculated. Using the standard deviation and mean on the 95% confidence interval (which 1.96) the upper and lower limit was obtained.

```
[1] "The upper limit = 101.107333457947"
[1] "The lower limit = 66.9447450343691"
```

Here we can notice that the upper and lower limit covers our net win of 82.19 value. Therefore, it means that the confidence interval contains our E(X) value.

(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.

A frequency distribution was constructed for Y which covers all the values of the outcome that can be obtained in the series. Which came out to be:

WW, WLW/LWW, LL, WLL/LWL

Where W = Win for Boston Red Sox & L = Win for New York Yankees.

Here taking out the profit/loss from the bets based on the game we get:

WW = 1000

WLW/LWW = 480

LL = -1040

WLL/LWL = -540

Then we get find the probability values for each outcome with their observed value which we can get from probability x 10000.

WW = 0.261 x 10000 = 2,610

WLW = 0.18502 x 10000 = 1850.2

LL = 0.231 x 10000 = 2,310

WLL = 0.13398 x 10000 = 1,339.8

LWL = 0.07938 x 10000 = 793.8

LWW = 0.10962 x 10000 = 1,096.2

Using the chi-square on the above observed values with their respective probability value we get

```
	Chi-squared test for given probabilities

data:  observed
X-squared = 1.1456, df = 3, p-value = 0.7661
```

Because frequency distributions of identical profit/loss values were summed (both their probability and their observed value) and displayed using the chi-square, the df is 3 and not 5.

(v) Use your observations of parts (ii) and (iii) above to describe whether your betting strategy is favorable to you.

The null hypothesis (Ho): states that the observed and experimental values acquired from the event are identical.

Alternative Hypothesis (H1): We found a discrepancy between observed and experimental values from the event.

Here we can notice that p-value is greater than 0.05(alpha value) therefore we don't have enough evidence to reject the null hypothesis meaning that observed and experimental values acquired from the event are identical.

Third task (Part Three)

Repeat part 1 above but now assume that the series is a best of five (5) series where the first team that wins three games wins the series with games alternating between Boston and New York, with the first game being played in New York.

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

Here we calculated the value of Probability by writing down the condition in which Red Sox will win the series which was:

WWW + WWLW + WLWW + LWWW +WLWLW +WLLWW +LLWWW + LWLWW + WWLLW + LWWLW

Then we calculated values for each and added them together to get the following probability of Red Sox winning the series.

```
[1] "The probability that Red Sox will win the series = 0.5036877"
```

Therefore, based on the above probability Red Sox has 50.36% of the winning the series if first game is played in New York, the second game is played in Boston, the third game is in New York, the fourth game is played in Boston, and the fifth game (if it becomes necessary) is in New York.

(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.

If the Red Sox win, we will collect $500, while if they lose, we will lose $520, according to our terms. So, for each match the probability of Red Sox winning and losing based on the home stadium of the team we get the net win and the standard deviation for the whole interaction

```
[1] "The expected net win = -39.8000000000002"
[1] "The standard deviation = 1131.09675978671"
```

Here the net win is negative -39.80 which is expected as three of the five games will be played in New York home stadium where the probability of Yankees winning is way more than Red Sox.

(iii) Use R to create 10,000 random values for X. Let these random values be denoted by Y. Use these Y values to estimate your expected net win by using a 95% confidence interval. Does this confidence interval contain E(X)?

Here using the rnorm function on the net win and standard deviation which we obtained in the previous task, a range of random 10000 numbers were generated and then their mean and standard is calculated. Using the standard deviation and mean on the 95% confidence interval (which 1.96) the upper and lower limit was obtained.

```
[1] "The upper limit = -3.58637691213594"
[1] "The lower limit = -47.6299078174198"
```

Here we can notice that the upper and lower limit covers our net win of -39.80 value. Therefore, it means that the confidence interval contains our E(X) value.

(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.

A frequency distribution was constructed for Y which covers all the values of the outcome that can be obtained in the series. Which came out to be:

WWW, WWLW / WLWW/ LWWW, WLWLW/ WLLWW/ LLWWW/ LWLWW/ WWLLW/ LWWLW, LLL, LLWL/ LWLL/ WLLL, LWLWL/ LWWLL/ WWLLL/ WLWLL/ LLWWL/ WLLWL

Where W = Win for Boston Red Sox & L = Win for New York Yankees.

Here taking out the profit/loss from the bets based on the game we get:

WWW = 1500

WWLW / WLWW/ LWWW = 980

WLWLW/ WLLWW/ LLWWW/ LWLWW/ WWLLW/ LWWLW = 460

LLL = -1560

LLWL/ LWLL/ WLLL = -1060

LWLWL/ LWWLL/ WWLLL/ WLWLL/ LLWWL/ WLLWL = -560

Then we get find the probability values for each outcome with their observed value which we can get from probability x 10000.

WWW = 0.11745 X 10000 = 1174.5

WWLW = 0.083259 x 10000 = 832.59

WLWW = 0.049329 x 10000 = 493.29

LWWW = 0.083259 x 10000 = 832.59

WLWLW = 0.01607445 x 10000 = 160.7445

WLLWW = 0.02713095 x 10000 = 271.3095

LLWWW = 0.02713095 x 10000 = 271.3095

LWLWW = 0.04579245 x 10000 = 457.9245

WWLLW = 0.02713095 x 10000 = 271.3095

LWWLW = 0.02713095 x 10000 = 271.3095

LLL = 0.12705 x 10000 = 1270.5

LLWL = 0.043659 x 10000 = 436.59

LWLL = 0.073689 x 10000 = 736.89

WLLL = 0.043659 x 10000 = 436.59

LWLWL = 0.05596855 x 10000 = 559.6855

LWWLL = 0.03316005 x 10000 = 331.6005

WWLLL = 0.03316005 x 10000 = 331.6005

WLWLL = 0.01964655 x 10000 = 196.4655

LLWWL = 0.03316005 x 10000 = 331.6005

WLLWL = 0.03316005 x 10000 = 331.6005

Using the chi-square on the above observed values with their respective probability value we get

```
        Chi-squared test for given probabilities

data:  observed
X-squared = 4.1159, df = 5, p-value = 0.5328
```

Because frequency distributions of identical profit/loss values were summed (both their probability and their observed value) and displayed using the chi-square, the df is 5 and not 20.

(v) Use your observations of parts (ii) and (iii) above to describe whether your betting strategy is favorable to you.

The null hypothesis (Ho): states that the observed and experimental values acquired from the event are identical.

Alternative Hypothesis (H1): We found a discrepancy between observed and experimental values from the event.

Here we can notice that p-value is greater than 0.05(alpha value) therefore we don’t have enough evidence to reject the null hypothesis meaning that observed and experimental values acquired from the event are identical.

Fourth task (Part Four)

Repeat part 1 above but now assume both teams will play the 2022 World Series. The series is a best of seven (7) series where the first team that wins four games wins the series. The team with home field advantage plays two games at home, three on the road(guest), and then if necessary two at home. Let's assume Boston Red Sox has the home field advantage against New York Yankees.

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

Here we calculated the value of Probability by writing down the condition in which Red Sox will win the series which was:

BBBB + NBBBB + BNBBB + BBNBB + BBBNB + NBBBNB + NBBNBB + NBNBBB + NNBBBB + BNBBNB + BNBNBB + BNNBBB + BBNBNB + BBNNBB + BBBNNB + NNNBBBB + NNBNBBB + NNBBNBB + NNBBBNB + NBNNBBB + NBNBNBB + NBNBBNB + NBBNNBB + NBBNBNB + NBBBNNB + BBBNNNB + BBNNNBB + BNNNBBB + BNNBNBB + BNNBBNB + BNBNBNB + BNBBNNB + BNBNNBB

Where B = Boston Red Sox winning the game & N = New York Yankees winning the game

Then we calculated values for each and added them together to get the following probability of Red Sox winning the series.

```
[1] "The probability that Red Sox will win the series because of their home
field advantage in the best of series = 0.623062170759999"
```

Therefore, based on the above probability Red Sox has 62.30% of the winning the series if first two games are played in Boston, the third, fourth and fifth games played in New York, the last two games (if it becomes necessary) is played in Boston.

(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.

If the Red Sox win, we will collect $500, while if they lose, we will lose $520, according to our terms. So, for each match the probability of Red Sox winning and losing based on the home stadium of the team we get the net win and the standard deviation for the whole interaction

```
[1] "The expected net win = 103.4"
[1] "The standard deviation = 1334.93298708212"
```

Here the net win is positive 103.4 which is expected as five of the seven games will be played in Boston home stadium where the probability of Red Sox winning is way more than Yankees.

(iii) Use R to create 10,000 random values for X. Let these random values be denoted by Y. Use these Y values to estimate your expected net win by using a 95% confidence interval. Does this confidence interval contain E(X)?

Here using the rnorm function on the net win and standard deviation which we obtained in the previous task, a range of random 10000 numbers were generated and then their mean and standard is calculated. Using the standard deviation and mean on the 95% confidence interval (which 1.96) the upper and lower limit was obtained.

```
[1] "The upper limit = 120.444224732901"
[1] "The lower limit = 67.9692889169899"
```

Here we can notice that the upper and lower limit covers our net win of 103.4 value. Therefore, it means that the confidence interval contains our E(X) value.

(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.

A frequency distribution was constructed for Y which covers all the values of the outcome that can be obtained in the series. Which came out to be:

BBBB, NBBBB / BNBBB / BBNBB / BBBNB, NBBBNB / NBBNBB / NBNBBB / NNBBBB / BNBBNB / BNBNBB / BNNBBB / BBNBNB / BBNNBB / BBBNNB, NNNBBBB / NNBNBBB / NNBBNBB / NNBBBNB / NBNNBBB / NBNBNBB / NBNBBNB / NBBNNBB / NBBNBNB / NBBBNNB / BBBNNNB / BBNNNBB / BNNNBBB / BNNBNBB / BNNBBNB / BNBNBNB / BNBBNNB / BNBNNBB, NNNN, BNNNN / NBNNN / NNBNN / NNNBN, BNNNBN / BNNBNN / BNBNNN / BBNNNN / NBNNBN / NBNBNN / NBBNNN / NNBNBN / NNBBNN / NNNBBN, BBBNNNN / BBNBNNN / BBNNBNN / BBNNNBN / BNBBNNN / BNBNBNN / BNBNNBN / BNNBBNN / BNNBNBN / BNNNBBN / NNNBBBN / NNBBBNN / NBBBNNN / NBBNBNN / NBBNNBN / NBNBNBN / NBNNBBN / NBNBBNN

Where B = Win for Boston Red Sox & N = Win for New York Yankees.

Here taking out the profit/loss from the bets based on the game we get:

BBBB = 2000

NBBBB / BNBBB / BBNBB / BBBNB = 1480

NBBBNB / NBBNBB / NBNBBB / NNBBBB / BNBBNB / BNBNBB / BNNBBB / BBNBNB / BBNNBB / BBBNNB = 960

NNNBBBB / NNBNBBB / NNBBNBB / NNBBBNB / NBNNBBB / NBNBNBB / NBNBBNB / NBBNNBB / NBBNBNB / NBBBNNB / BBBNNNB / BBNNNBB / BNNNBBB / BNNBNBB / BNNBBNB / BNBNBNB / BNBBNNB / BNBNNBB = 440

NNNN = -2080

BNNNN / NBNNN / NNBNN / NNNBN = -1580

BNNNBN / BNNBNN / BNBNNN / BBNNNN / NBNNBN / NBNBNN / NBBNNN / NNBNBN / NNBBNN / NNNBBN = -1080

BBBNNNN / BBNBNNN / BBNNBNN / BBNNNBN / BNBBNNN / BNBNBNN / BNBNNBN / BNNBBNN / BNNBNBN / BNNNBBN / NNNBBBN / NNBBBNN / NBBBNNN / NBBNBNN / NBBNNBN / NBNBNBN / NBNNBBN / NBNBBNN = -580

Then we get find the probability values for each outcome with their observed value which we can get from probability x 10000.

Boston winning the series

BBBB = 0.068121 * 10000 = 681.21

NBBBB = 0.02219805 * 10000 = 221.9805

BNBBB = 0.02219805 * 10000 = 221.9805

BBNBB = 0.03746655 * 10000 = 374.6655

BBBNB = 0.03746655 * 10000 = 374.6655

NBBBNB = 0.01573595 * 10000 = 157.3595

NBBNBB = 0.01573595 * 10000 = 157.3595

NBNBBB = 0.01573595 * 10000 = 157.3595

NNBBBB = 0.009323181 * 10000 = 93.23181

BNBBNB = 0.02713095 * 10000 = 271.3095

BNBNBB = 0.02713095 * 10000 = 271.3095

BNNBBB = 0.02713095 * 10000 = 271.3095

BBNBNB = 0.04579245 * 10000 = 457.9245

BBNNBB = 0.04579245 * 10000 = 457.9245

BBBNNB = 0.04579245 * 10000 = 457.9245

NNNBBBB = 0.006609099 * 10000 = 66.09099

NNBNBBB = 0.006609099 * 10000 = 66.09099

NNBBNBB = 0.006609099 * 10000 = 66.09099

NNBBBNB = 0.003915736 * 10000 = 39.15736

NBNNBBB = 0.01115504 * 10000 = 111.5504

NBNBNBB = 0.01115504 * 10000 = 111.5504

NBNBBNB = 0.006609099 * 10000 = 66.09099

NBBNNBB = 0.01115504 * 10000 = 111.5504

NBBNBNB = 0.006609099 * 10000 = 66.09099

NBBBNNB = 0.006609099 * 10000 = 66.09099

BBBNNNB = 0.01115504 * 10000 = 111.5504

BBNNNBB = 0.01882782 * 10000 = 188.2782

BNNNBBB = 0.01115504 * 10000 = 111.5504

BNNBNBB = 0.01115504 * 10000 = 111.5504

BNNBBNB = 0.006609099 * 10000 = 66.09099

BNBNBNB = 0.006609099 * 10000 = 66.09099

BNBBNNB = 0.006609099 * 10000 = 66.09099

BNBNNBB = 0.01115504 * 10000 = 111.5504

New York winning the series

NNNN = 0.053361 * 10000 = 533.61

BNNNN = 0.04052895 * 10000 = 405.2895

NBNNN = 0.04052895 * 10000 = 405.2895

NNBNN = 0.02401245 * 10000 = 240.1245

NNNBN = 0.02401245 * 10000 = 240.1245

BNNNBN = 0.01392722 * 10000 = 139.2722

BNNBNN = 0.01392722 * 10000 = 139.2722

BNBNNN = 0.01392722 * 10000 = 139.2722

BBNNNN = 0.02350679 * 10000 = 235.0679

NBNNBN = 0.01392722 * 10000 = 139.2722

NBNBNN = 0.01392722 * 10000 = 139.2722

NBBNNN = 0.01392722 * 10000 = 139.2722

NNBNBN = 0.008251551 * 10000 = 82.51551

NNBBNN = 0.008251551 * 10000 = 82.51551

NNNBBN = 0.008251551 * 10000 = 82.51551

BBBNNNN = 0.008077788 * 10000 = 80.77788

BBNBNNN = 0.008077788 * 10000 = 80.77788

BBNNBNN = 0.008077788 * 10000 = 80.77788

BBNNNBN = 0.01363394 * 10000 = 136.3394

BNBBNNN = 0.0047859 * 10000 = 47.859

BNBNBNN = 0.0047859 * 10000 = 47.859

BNBNNBN = 0.008077788 * 10000 = 80.77788

BNNBBNN = 0.0047859 * 10000 = 47.859

BNNBNBN = 0.008077788 * 10000 = 80.77788

BNNNBBN = 0.008077788 * 10000 = 80.77788

NNNBBBN = 0.0047859 * 10000 = 47.859

NNBBBNN = 0.002835533 * 10000 = 28.35533

NBBBNNN = 0.0047859 * 10000 = 47.859

NBBNBNN = 0.0047859 * 10000 = 47.859

NBBNNBN = 0.008077788 * 10000 = 80.77788

NBNBNBN = 0.008077788 * 10000 = 80.77788

NBNNBBN = 0.008077788 * 10000 = 80.77788

NBNBBNN = 0.0047859 * 10000 = 47.859

Using the chi-square on the above observed values with their respective probability value we get

```
	Chi-squared test for given probabilities

data:  observed
X-squared = 2894.8, df = 7, p-value < 2.2e-16
```

Because frequency distributions of identical profit/loss values were summed (both their probability and their observed value) and displayed using the chi-square, the df is 7 and not 66.

(v) Use your observations of parts (ii) and (iii) above to describe whether your betting strategy is favorable to you.

The null hypothesis (Ho): states that the observed and experimental values acquired from the event are identical.

Alternative Hypothesis (H1): We found a discrepancy between observed and experimental values from the event.

Here we can notice that p-value will be less than 0.05(alpha value) therefore we reject the null hypothesis meaning that we found a discrepancy between observed and experimental values from the event.

CONCLUSION

Boston Red Sox has a greater chance of winning the best of three series when they play the opening games in their home stadium as the last game if played will be played in their home stadium as well. In addition, the best of five series has a little better odd of winning for Boston Red Sox even if most three of the five games will be played in New York home stadium. But overall, it will be net negative for us as most of the games will be New York home stadium where they have home turf advantage. Lastly in the best of seven games series, Boston Red Sox has a better advantage of winning the whole series as majority of the games will be played in Boston if it goes to a seven-game series. So, my recommendation we would to bet on Boston Red Sox when majority of their games are being played on their home ground as most of the time, they will win the series. Finally, as a result of this assignment, I learned how to generate random variables and utilize the chi-square goodness of fit test.

REFERENCES

1) Z. (2020, October 21). How to Perform a Chi-Square Goodness of Fit Test in R. Statology. https://www.statology.org/chi-square-goodness-of-fit-test-in-r/

2) RPubs - Chi-Square Goodness of Fit Test. (2020, April 4). Priyank Goyal. https://rpubs.com/pg2000in/ChiSquareGoodnessFit

3) Datamentor. (2017, November 27). R Program to Generate Random Number from Standard Distributions. https://www.datamentor.io/r-programming/examples/random-number/#:%7E:text=Random%20numbers%20from%20a%20normal,is%20between%200%20and%201%20.