Module-1: Analysis of a Betting Strategy in Sports

By: Nithin Reddy.P (NEU ID:002896440)

Subject:ALY6050: Introduction to Enterprise Analytics
Under the guidance of Shahram Sattar
Submission Date:25-Feb-2024

Introduction:

In this assignment, we have two teams Boston Red Sox and Yankees playing a series, in a home as well away games. The winner should win two out of three games or three out of five games. The main aim is to apply probability theory for a betting simulation.

Problem:

Teams from NY and Boston are playing a three match series and the first team that wins two games will win the series. 0.6 is the probability of Red winning a game at home and for Yankees its 0.57 at their home turf. If a bet is placed you would win $500 if Red wins and loose $520 in the other case.

Facts:

1. 0.57 is the probability of N.K.Y wining at home and for an away game it is 0.43
2. For Sox 0.6 is probability winning a home game and 0.4 for a touring game
3. Gains if Red Sox wins is $500
4. Loss if Red Sox lost is $520

Part-1:

Scenario-1: First game is played in Boston, second in New York and the last is at Boston.

For B.R.S to Win the series:

It can be calculated by adding the probability of Red Sox fining the first two games, wins in first and last game as well as wins in second and last game. From Fig-1, we could see that the probability of wining the series is 0.5664 (sum of respective probabilities are 0.258, 0.2052 and 0.1032).

Probability distribution of Net win and calculation of expected net win and variance:

We are aware that $500 is the gain if Red Sox wins and a loose $520 is case of a series loss. We have identified the 0. 5664 as probability of wining the series, hence the probability of the loss would be 0.4336 (1-probability of wining the series). From Fig-2, we could see that the respective values and probabilities are assigned to vectors and expected net win is calculated using formula $E(X)=\sum ixi \cdot P(X=xi)$ I,e 57.728. The variance is calculated using formula $Var(X)=\sum i=1n(xi-\mu)2 \cdot pi$ I,e 255512.9, the standard deviation is calculated using formula Sqrt(Var(X)), ie 505.4829.

Estimating Expected Net win using 95% confidence interval:

Please refer to Fig-3, a random sample is generated using rnorm function with the available mean and standard deviation. The sample mean, standard error, and z score for 95% interval have be calculated. By sung the sample

mean, z score and standard error we have calculated the upper and lower bounds of the confidence interval have been found and we could see that the confidence interval (48.95895, 68.72811) contains the expected net win (57.728).

Frequency Distribution and Chi-Square Goodness of fit:
Please refer to Fig-4, the generated random values haven been divided into bins and counting the number of frequencies in each bin. Later, the expected frequencies are calculated using cumulative distribution function. Finally, chi-square goodness of fit is calculated for distributions of observed and expected frequencies. The hypothesis is,

H0= There is no significant difference between the observed frequencies and the expected frequencies

H1= There is significant difference between the observed frequencies and the expected frequencies

Form the output, we could clearly see that the p-value is greater than significance level (0.05), hence we fail to reject the null hypothesis, concluding that observed frequencies are consistent with expected frequencies.

Analysis of betting strategy of Scenario-1:
Based on the analysis the betting strategy is favourable and the below facts support the strategy, suggesting potential profitability in betting on Red Sox:

1. Probability of Red Sox winning the 3-match series is 0.5664
2. The obtained net win by a series win is $57.728 with a SD of 505.4829
3. The confidence interval of (48.95895, 68.72811) contains the expected net win (57.728).
4. The Chi-square good ness-of-fit suggest that observed frequencies closely match with expected frequencies, indicating a reliable strategy

Part-2:

Scenario-2: First game is played in New York, second in Boston and the last is at New York.
Please look at Figures-6,7, 8 and 9, the same steps have been performed like scenario-1 and please refer to the below findings.

Analysis of betting strategy of Scenario-2:
Findings:

1. The probability of Red Sox wining is 0.47902
2. The obtained expected net win is $-31.3996 with standard deviation of $509.5508
3. The confidence interval (-50.1049, -30.06679) contains expected net win ($81.75)

4. The Chi-square good ness of fit suggest that observed frequencies closely match with expected frequencies, indicating a reliable strategy

The probability of Red Sox winning and expected net win changed from scenatrio-1 to scenario-2 but based on the above findings we can conclude that the betting strategy is not favourable, suggesting a loss of $-31.3.

Part-3:

Scenario-3: If a series is a best-of-five series the team that wins 3 games take the series, the games alter between Boston and NY, the first is at Boston. Please refer to Figures-10,-11, -12 and -13, the same steps have been performed similar to scenario-1 and kindly refer to the below findings.

1. The probability of Red Sox wining is 0.48174
2. The obtained expected net loss is $-28.6252 with standard deviation of $ 501.6774
3. The confidence interval (-32.98235, 13.34523) contains expected new win -28.6252
4. The Chi-square good ness of fit suggest that observed frequencies closely match with expected frequencies, indicating a reliable strategy

The probability of Red Sox winning is less than 50% and we would incur a loss of $-28.62, based on the observations we conclude that the betting strategy is not favourable as the outcome suggests a loss.

Conclusion:

In conclusion, the betting strategy was examined under three scenarios. In the first scenario, the strategy was favourable. The probability of Red Sox winning as high, the expected net win as high and the chi-squared good ness of fit suggested the observed and expected frequencies closely match, which validates the strategy. On the other hand, the second and third scenario indicated that the strategy was not favourable as the net win as negative, indications a loss. Hence, it is crucial to consider and monitor the order of games and format of the series to formulate a better betting strategy.

Appendix:

Scenario-1

Fig-1: Red Sox wins the series

```
> # Probability of Red Sox winning a game in Boston
> p_red_sox_boston <- 0.6
>
> # Probability of Red Sox winning a game in New York
> p_red_sox_win_ny <- 1 - 0.57
>
> # Probability of Yankees winning a game in New York
> p_yankees_new_york <- 0.57
>
> # Probability of Yankees winning a game in Boston
> p_yankees_boston <- 1-0.6
>
> # No of games
>
> no_games <- 3
> # Probability of Red Sox winning the first two games
>
> p_red_sox_win_1st_2nd <- p_red_sox_boston * p_red_sox_win_ny
> p_red_sox_win_1st_2nd
[1] 0.258
>
> # Probability of Red Sox winning the first game, Yankees winning the second, and Red Sox winning the third
>
> p_red_sox_win_1st_3rd <- p_red_sox_boston * p_yankees_new_york * p_red_sox_boston
> p_red_sox_win_1st_3rd
[1] 0.2052
>
> # Probability of Red Sox winning the second and third games
>
> p_red_sox_win_2nd_3rd <- p_yankees_boston*p_red_sox_win_ny * p_red_sox_boston
> p_red_sox_win_2nd_3rd
[1] 0.1032
>
> # Total probability of Red Sox winning the series
>
> p_red_sox_win_series <- p_red_sox_win_1st_2nd + p_red_sox_win_1st_3rd + p_red_sox_win_2nd_3rd
>
> cat("Probability of Red Sox winning the series is:", p_red_sox_win_series)
Probability of Red Sox winning the series is: 0.5664
```

Fig-2: Net win and Variance

```
> net_win_red_sox_win_series <- 500
> net_win_red_sox_lose_series <- -520
>
> net_win_distribution <- c(net_win_red_sox_win_series, net_win_red_sox_lose_series)
> probabilities <- c(p_red_sox_win_series, 1 - p_red_sox_win_series)
>
> expected_net_win <- (net_win_red_sox_win_series*p_red_sox_win_series) +
+   (net_win_red_sox_lose_series*(1-p_red_sox_win_series))
>
> cat("Expected net win of winnng the series is:", expected_net_win)
Expected net win of winnng the series is: 57.728>
> variance <- sum(probabilities*(net_win_distribution-expected_net_win )^2)
> cat("Variance of winnng the series is:", variance)
Variance of winnng the series is: 255512.9>
> standard_dev <- sqrt(variance)
> cat("Standard Deviation of winnng the series is:", standard_dev )
Standard Deviation of winnng the series is: 505.4829
```

Fig-3: Estimating net win using 95% confidence interval:

```
> Y <- rnorm(10000, mean = expected_net_win, sd = standard_dev )
>
> # sample mean (Y_bar)
> Y_bar <- mean(Y)
> Y_bar
[1] 58.84353
> # standard error (SE)
> SE <- sd(Y) / sqrt(length(Y))
> SE
[1] 5.043245
> # z-score for 95% confidence interval
> z <- qnorm(0.975)
> z
[1] 1.959964
> # Calculate lower and upper bounds of the confidence interval
> lower_bound <- Y_bar - z * SE
> lower_bound
[1] 48.95895
> upper_bound <- Y_bar + z * SE
> upper_bound
[1] 68.72811
> # Check if expected_net_win falls within the confidence interval
> if (expected_net_win >= lower_bound && expected_net_win <= upper_bound) {
+   cat("The confidence interval contains the expected net win.")
+ } else {
+   cat("The confidence interval does not contain the expected net win.")
+ }
The confidence interval contains the expected net win.
```

Fig-4: Frequency dist and Chi-Square Goodness of fit:

```
> # number of bins for the frequency distribution
> num_bins <- 10
>
> # Creating bins for the frequency distribution
> breaks <- seq(min(Y), max(Y), length.out = num_bins + 1)
>
> # observed frequencies for each bin
> observed_freq <- table(cut(Y, breaks = breaks, include.lowest = TRUE))
> observed_freq

[-1.83e+03,-1.45e+03] (-1.45e+03,-1.06e+03]      (-1.06e+03,-682]           (-682,-300]
                   18                    108                   572                  1683
          (-300,81.7]            (81.7,464]             (464,846]        (846,1.23e+03]
                 2824                   2688                  1543                   459
  (1.23e+03,1.61e+03]   (1.61e+03,1.99e+03]
                   90                    15
> # expected frequencies based on the distribution of X
> expected_freq <- diff(pnorm(breaks, mean = expected_net_win, sd = standard_dev)) * length(Y)
> expected_freq
 [1]   13.70030  117.68423  584.05755 1678.02156 2794.96333 2700.96600 1514.27269  492.10933   92.56066
[10]   10.05593
> # Perform the chi-squared goodness-of-fit test
> chisq_test <- chisq.test(x=observed_freq, p = expected_freq/sum(expected_freq))
> chisq_test

        Chi-squared test for given probabilities

data:  observed_freq
X-squared = 8.0466, df = 9, p-value = 0.5295

>
> if(chisq_test$p.value<=0.05){
+   cat("p-value is less than 0.05. Therefore, we reject the null hypothesis ")
+ } else {
+   cat("p-value is greater than 0.05. Therefore, we fail to reject the null hypothesis ")
+ }
p-value is greater than 0.05. Therefore, we fail to reject the null hypothesis
```

Scenario-2

Fig-6: Probability of R.S winning the series

```
> #1: Calculate the probability that the Red Sox will win the series
>
> # Probability of Red Sox winning the first two games
>
> p2_red_sox_win_1st_2nd <- p_red_sox_win_ny*p_red_sox_boston
> p2_red_sox_win_1st_2nd
[1] 0.258
>
> # Probability of Red Sox winning the first game, Yankees winning the second, and Red Sox winning the third
>
> p2_red_sox_win_1st_3rd <- p_red_sox_win_ny * p_yankees_boston  * p_red_sox_win_ny
> p2_red_sox_win_1st_3rd
[1] 0.07396
>
> # Probability of Red Sox winning the second and third games
>
> p2_red_sox_win_2nd_3rd <-  p_yankees_new_york*p_red_sox_boston*p_red_sox_win_ny
> p2_red_sox_win_2nd_3rd
[1] 0.14706
>
> # Total probability of Red Sox winning the series
>
> p2_red_sox_win_series <- p2_red_sox_win_1st_2nd + p2_red_sox_win_1st_3rd + p2_red_sox_win_2nd_3rd
>
> cat("Probability of Red Sox winning the series is:", p2_red_sox_win_series)
Probability of Red Sox winning the series is: 0.47902
```

Fig-7: Expected Net win and Variance

```
> #2: 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.
>
> #Win of each out come
>
> net_win_red_sox_win_series <- 500
> net_win_red_sox_lose_series <- -520
>
> net_win_dist <- c(net_win_red_sox_win_series, net_win_red_sox_lose_series)
> prob <- c(p2_red_sox_win_series, 1 - p2_red_sox_win_series)
>
> exp_net_win <- (net_win_red_sox_win_series*p2_red_sox_win_series) +
+   (net_win_red_sox_lose_series*(1-p2_red_sox_win_series))
>
> cat("Expected net win of winnng the series is:", exp_net_win)
Expected net win of winnng the series is: -31.3996>
> var <- sum(prob*(net_win_dist-exp_net_win )^2)
> cat("Variance of winnng the series is:", var)
Variance of winnng the series is: 259642.1>
> st_dev <- sqrt(var)
> cat("Standard Deviation of winnng the series is:", st_dev )
Standard Deviation of winnng the series is: 509.5508
```

Fig-8: Estimating Expected Net win using 95% confidence interval:

```
> # Generate 10,000 random values for X
> Y2 <- rnorm(10000, mean = exp_net_win, sd = st_dev )
>
> # sample mean (Y_bar)
> Y2_bar <- mean(Y2)
> Y2_bar
[1] -40.08584
> # standard error (SE)
> SE2 <- sd(Y2) / sqrt(length(Y2))
> SE2
[1] 5.111856
> # z-score for 95% confidence interval
> z2 <- qnorm(0.975)
> z2
[1] 1.959964
> # Calculate lower and upper bounds of the confidence interval
> lower_bound2 <- Y2_bar - z2 * SE2
> lower_bound2
[1] -50.1049
> upper_bound2 <- Y2_bar + z2 * SE2
> upper_bound2
[1] -30.06679
>
> # Check if expected_net_win falls within the confidence interval
> if (exp_net_win >= lower_bound2 && exp_net_win <= upper_bound2) {
+   cat("The confidence interval contains the expected net win.")
+ } else {
+   cat("The confidence interval does not contain the expected net win.")
+ }
The confidence interval contains the expected net win.
```

Fig-9: Frequency Distribution and Chi-Square Goodness of fit:

```
> # number of bins for the frequency distribution
> num_bins2 <- 10
>
> # Creating bins for the frequency distribution
> breaks2 <- seq(min(Y2), max(Y2), length.out = num_bins2 + 1)
>
> # observed frequencies for each bin
> observed_freq2 <- table(cut(Y2, breaks = breaks2, include.lowest = TRUE))
> observed_freq2

[-1.96e+03,-1.58e+03]  (-1.58e+03,-1.2e+03]       (-1.2e+03,-815]           (-815,-432]
                   15                    108                   526                  1594
           (-432,-49]             (-49,334]             (334,717]         (717,1.1e+03]
                 2690                  2718                  1648                   588
   (1.1e+03,1.48e+03]   (1.48e+03,1.87e+03]
                  106                     7
> # expected frequencies based on the distribution of X
> expected_freq2 <- diff(pnorm(breaks2, mean = exp_net_win, sd = st_dev)) * length(Y2)
> expected_freq2
 [1]   11.02940   98.41468  510.03289 1538.18843 2703.52206 2771.46263 1657.15325  577.52465  117.14399
[10]   13.80332
> # Perform the chi-squared goodness-of-fit test
> chisq_test2 <- chisq.test(x=observed_freq2, p = expected_freq2/sum(expected_freq2))
> chisq_test2

        Chi-squared test for given probabilities

data:  observed_freq2
X-squared = 10.639, df = 9, p-value = 0.3013

>
> if(chisq_test2$p.value<=0.05){
+   cat("p-value is less than 0.05. Therefore, we reject the null hypothesis ")
+ } else {
+   cat("p-value is greater than 0.05. Therefore, we fail to reject the null hypothesis ")
+ }
p-value is greater than 0.05. Therefore, we fail to reject the null hypothesis
`
```

Scenario-3:

Fig-10: Probability of Red Sox winning the series

```
> #1: Calculate the probability that the Red Sox will win the series
>
> # Probability of Red Sox winning two games in Boston
>
> p_red_sox_win_Boston <- (p_red_sox_boston)^2*p_red_sox_win_ny
> p_red_sox_win_Boston
[1] 0.1548
>
> # Probability of Red Sox winning two games in NY
>
> p_red_sox_win_NY <- (p_red_sox_win_ny)^2 * p_red_sox_boston
> p_red_sox_win_NY
[1] 0.11094
>
> # Probability of Red Sox winning all three in Boston
>
> p_red_sox_win_3_in_Boston <- (p_red_sox_boston)^3
> p_red_sox_win_3_in_Boston
[1] 0.216
>
> # Probability that the Red Sox will win the series
>
> total_prob <- p_red_sox_win_Boston + p_red_sox_win_NY + p_red_sox_win_3_in_Boston
>
> cat("Probability that the Red Sox will win the series is:", total_prob )
Probability that the Red Sox will win the series is: 0.48174>
```

Fig-11: Expected Net win and Variance

```
> #2: 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.
>
> #Win of each out come
>
> net_win_red_sox_win_series <- 500
> net_win_red_sox_lose_series <- -520
>
> net_win_dist2 <- c(net_win_red_sox_win_series, net_win_red_sox_lose_series)
> prob2 <- c(total_prob, 1 - total_prob)
>
> exp_net_win2 <- (net_win_red_sox_win_series*total_prob) +
+   (net_win_red_sox_lose_series*(1-total_prob))
>
> cat("Expected net win of winnng the series is:", exp_net_win2)
Expected net win of winnng the series is: -28.6252>
> var2 <- sum(prob*(net_win_dist-exp_net_win )^2)
> cat("Variance of winnng the series is:", var2)
Variance of winnng the series is: 251680.2>
> st_dev2 <- sqrt(var2)
> cat("Standard Deviation of winnng the series is:", st_dev2 )
Standard Deviation of winnng the series is: 501.6774>
```

Fig-12: Estimating Expected Net win using 95% confidence interval:

```
> ##3.  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)?
>
> # Generate 10,000 random values for X
> Y3 <- rnorm(10000, mean = exp_net_win2, sd = st_dev2 )
>
> # sample mean (Y_bar)
> Y3_bar <- mean(Y3)
> Y3_bar
[1] -23.16379
> # standard error (SE)
> SE3 <- sd(Y3) / sqrt(length(Y3))
> SE3
[1] 5.00956
> # z-score for 95% confidence interval
> z3 <- qnorm(0.975)
> z3
[1] 1.959964
> # Calculate lower and upper bounds of the confidence interval
> lower_bound3 <- Y3_bar - z3 * SE3
> lower_bound3
[1] -32.98235
> upper_bound3 <- Y3_bar + z3 * SE3
> upper_bound3
[1] -13.34523
```

Fig-13: Frequency Distribution and Chi-Square Goodness of fit:

```
> #H0= there is no significant difference between the observed frequencies and the expected frequencies
> #H1=  there is significant difference between the observed frequencies and the expected frequencies
>
> # number of bins for the frequency distribution
> num_bins3 <- 10
>
> # Creating bins for the frequency distribution
> breaks3 <- seq(min(Y3), max(Y3), length.out = num_bins3 + 1)
>
> # observed frequencies for each bin
> observed_freq3 <- table(cut(Y3, breaks = breaks3, include.lowest = TRUE))
> observed_freq3

[-1.76e+03,-1.4e+03] (-1.4e+03,-1.04e+03]     (-1.04e+03,-687]          (-687,-331]
                  27                   168                   759                  1711
         (-331,24.5]           (24.5,380]            (380,736]       (736,1.09e+03]
                2697                  2486                  1516                   520
 (1.09e+03,1.45e+03]   (1.45e+03,1.8e+03]
                 108                     8
> # expected frequencies based on the distribution of X
> expected_freq3 <- diff(pnorm(breaks3, mean = exp_net_win2, sd = st_dev2)) * length(Y2)
> expected_freq3
 [1]   28.55445  184.02526  730.00843 1784.67656 2691.23121 2504.29550 1437.90654  509.14515  111.06788
[10]   14.90733
> # Perform the chi-squared goodness-of-fit test
> chisq_test3 <- chisq.test(x=observed_freq3, p = expected_freq3/sum(expected_freq3))
> chisq_test3

        Chi-squared test for given probabilities

data:  observed_freq3
X-squared = 13.57, df = 9, p-value = 0.1385
```

References:

Probability course: Expected Value and Variance; [online]: https://www.probabilitycourse.com/chapter4/4_1_2_expected_val_variance.php

Open Stax: Expected Value and Standard Deviation; [online]: https://openstax.org/books/statistics/pages/4-2-mean-or-expected-value-and-standard-deviation#:~:text=To%20find%20the%20expected%20value,%E2%88%91%20x%20P%20(%20x%20)%20.