Northeastern University
College of Professional Studies

Module 1-Project: Analysis of a Betting Strategy in Sports

Thuy Nhu Thao Tran

COLLEGE OF PROFESSIONAL STUDIES,

NORTHEASTERN UNIVERSITY

ALY6050 - Introduction to Enterprise Analytics

Instructor: Adam Jones

February 23, 2025

Introduction

This paper examines a betting strategy for a baseball series between the Boston Red Sox and the New York Yankees, comparing various game sequences and series durations. The purpose is to assess if betting on the Red Sox is advantageous by comparing theoretical probability to simulated results.

This report includes three scenarios:

- Part 1: Best-of-three series (Boston, New York, Boston).
- Part 2: Best-of-three series (New York, Boston, New York).
- Part 3: Best-of-five series.

Each of these parts will involve probabilities, expected winnings, simulations, and Chi-squared tests. The findings demonstrate the importance of both home-field advantage and series format in affecting betting profitability and risk. Although some strategies appear favorable, their high outcome variance indicates considerable risk. This document is a report to assist bettors in making data-based decisions grounded in statistical models.

Analysis

Part 1: Games in Boston, New York, Boston

1. **Step 1: Calculate the probability that the Red Sox will win the series.**

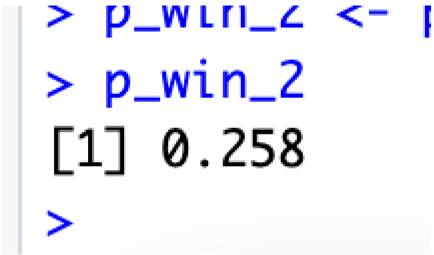
I first compute the total probability of the Red Sox winning the series by evaluating two scenarios:

- Winning in 2 games: The Red Sox win in both the first and second games.
- Winning in 3 games: The Red Sox secure victory in one game, lose the other, and then win in the third game.

The computations are as follows:

- P(Win in 2 games) = P(Win at home) X P(Win away) = 0.6 x 0.43 = 0.258

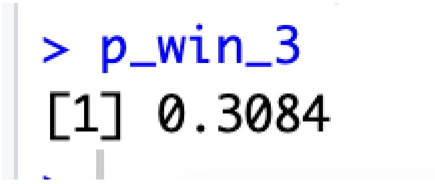
Output:



```
> p_win_2 <-
> p_win_2
[1] 0.258
>
```

- P(Win in 3 games) = (0.6 × 0.57 × 0.6) + (0.4 × 0.43 × 0.6) = 0.3084

Output:

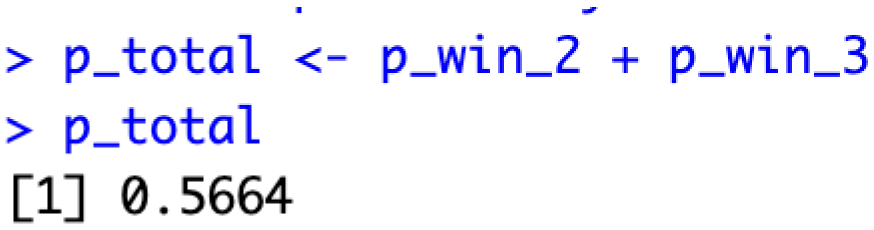


```
> p_win_3
[1] 0.3084
```

Total probability:

P(Total) = P(Win in 2 games) + P(Win in 3 games) = 0.258 + 0.3084 = 0.5664

Output:



```
> p_total <- p_win_2 + p_win_3
> p_total
[1] 0.5664
```

Interpretation:

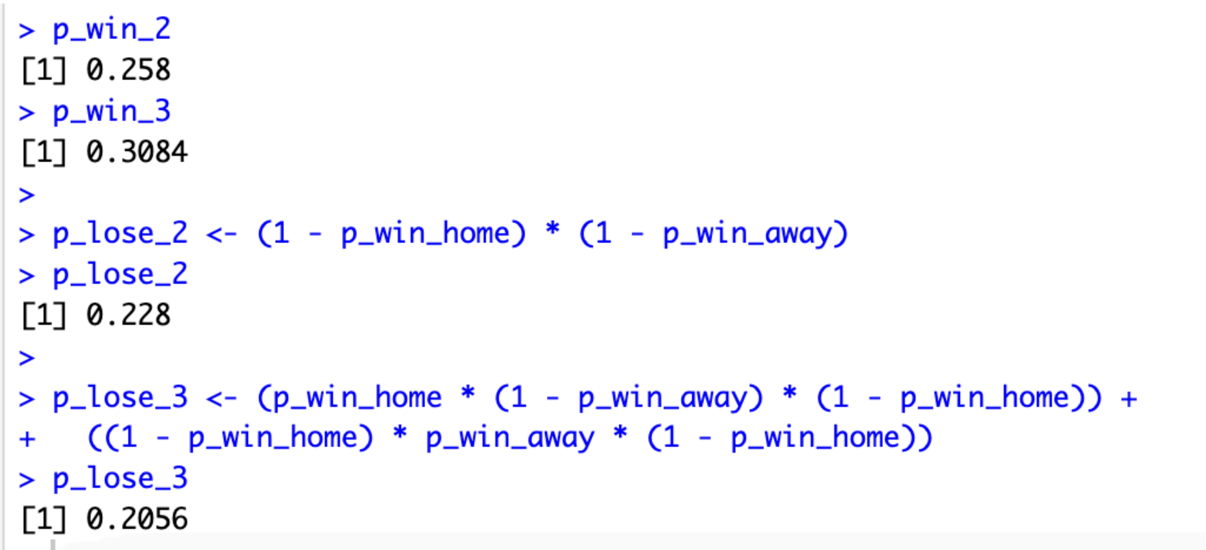
The probability of the Red Sox winning the series under the above conditions is 56.64%.

2. **Step 2: Construct the theoretical probability distribution for net winnings (X).**

Next, I formulate the theoretical distribution of net wins X for each potential outcome:

Outcome	Net Winnings ($)	Probability
Win in 2 games	+1000	0.258
Win in 3 games	+480	0.3084
Lose in 2 games	-1040	0.228
Lose in 3 games	-540	0.2056

Output:

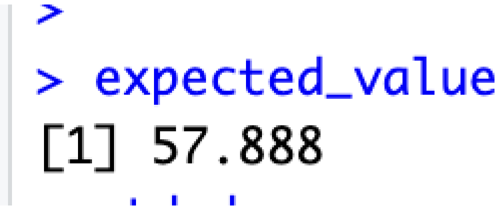


```
> p_win_2
[1] 0.258
> p_win_3
[1] 0.3084
>
> p_lose_2 <- (1 - p_win_home) * (1 - p_win_away)
> p_lose_2
[1] 0.228
>
> p_lose_3 <- (p_win_home * (1 - p_win_away) * (1 - p_win_home)) +
+   ((1 - p_win_home) * p_win_away * (1 - p_win_home))
> p_lose_3
[1] 0.2056
```

Expected Value (Mean of X):

E(X) = (1000 × 0.258) + (480× 0.3084) + (−1040 × 0.228) + (−540 × 0.2056) = 57.89

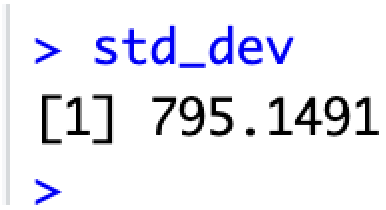
Output:



```
>
> expected_value
[1] 57.888
```

Standard Deviation: 795.15

Output:



```
> std_dev
[1] 795.1491
>
```

Interpretation:

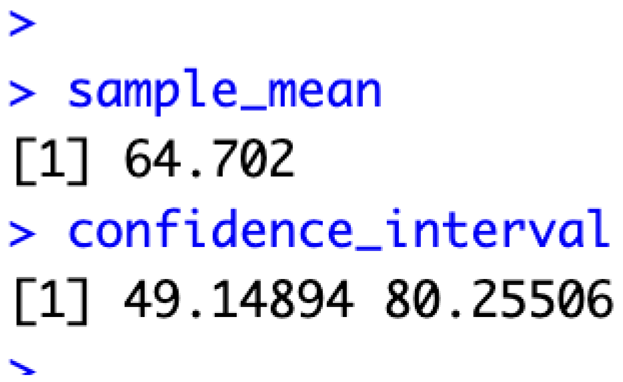
The expected net profit per series is $57.89, accompanied by a standard deviation of $795.15, indicating considerable variety in potential results.

3. **Step 3: Simulate 10,000 series and compute a 95% confidence interval.**

I simulated a 10,000 series, randomly producing outcomes according to the specified probabilities. The outcomes were:

- Sample Mean: $64.70
- 95% Confidence Interval: [$49.15, $80.26]

Output:



```
>
> sample_mean
[1] 64.702
> confidence_interval
[1] 49.14894 80.25506
>
```

Interpretation:

The confidence interval encompasses the theoretical anticipated value ($57.89), indicating that the simulation corresponds with theoretical expectations.

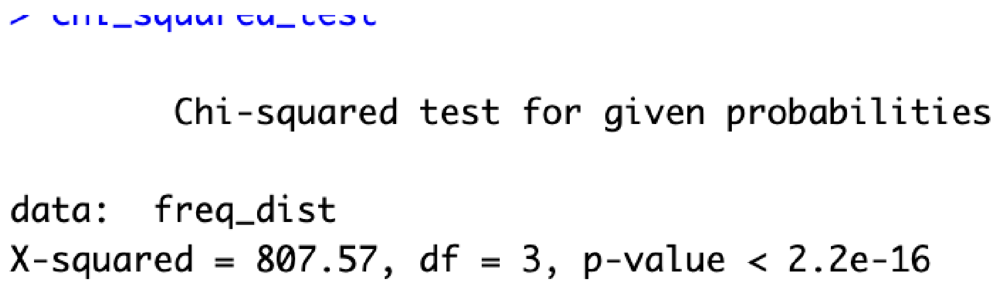
4. **Step 4: Frequency distribution and Chi-squared test.**

I evaluated the simulated frequency distribution with the theoretical distribution by using a Chi-squared test.

- Chi-squared Statistic: 807.57

- p-value: < 2.2e-16

Output:



```
> chi_squared_test

	Chi-squared test for given probabilities

data:  freq_dist
X-squared = 807.57, df = 3, p-value < 2.2e-16
```

Interpretation:

The substantial p-value indicates considerable difference between the simulated and theoretical distributions. The distinction may result from the random nature of simulations or foundational assumptions.

5. **Step 5: Determine if my betting strategy is favorable.**

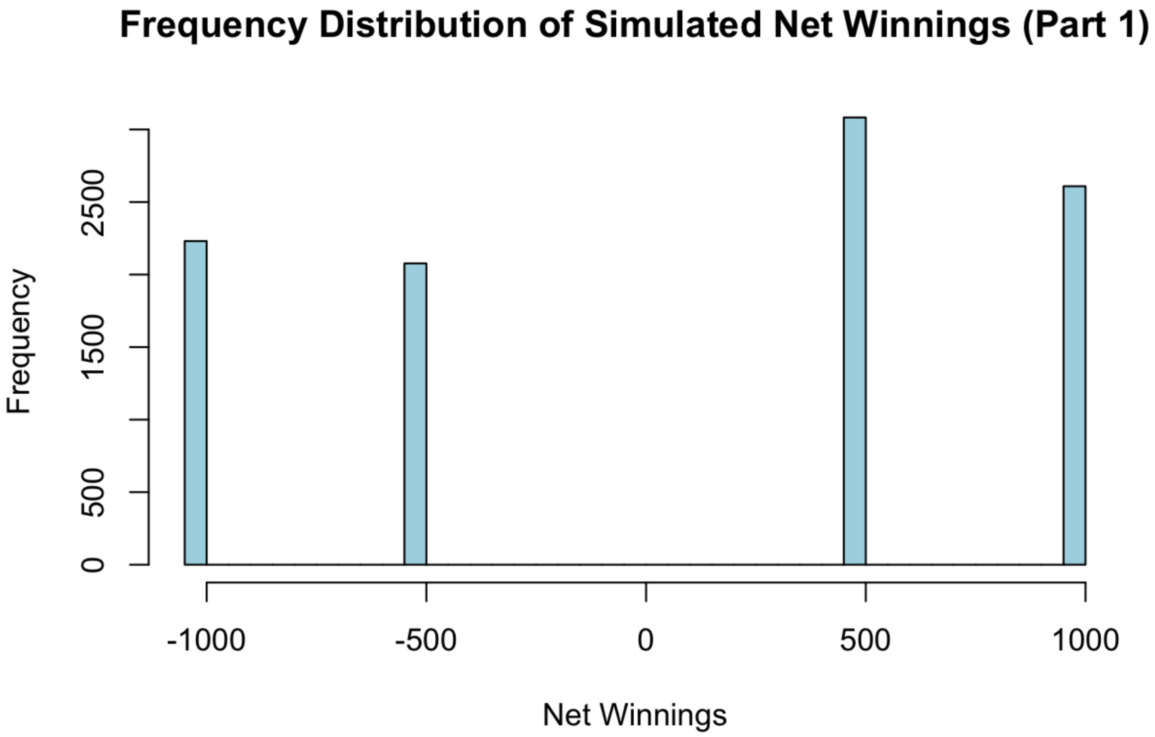
My observation through steps 2 and 3:

- Step 2: Simulation results validate this anticipation, presenting a 95% confidence interval of [$49.15, $80.26], which includes the theoretical mean.
- Step 3: The Chi-squared test shows that the theoretical and simulated distributions are not the same. This is likely because of sample variability.

Final Interpretation:

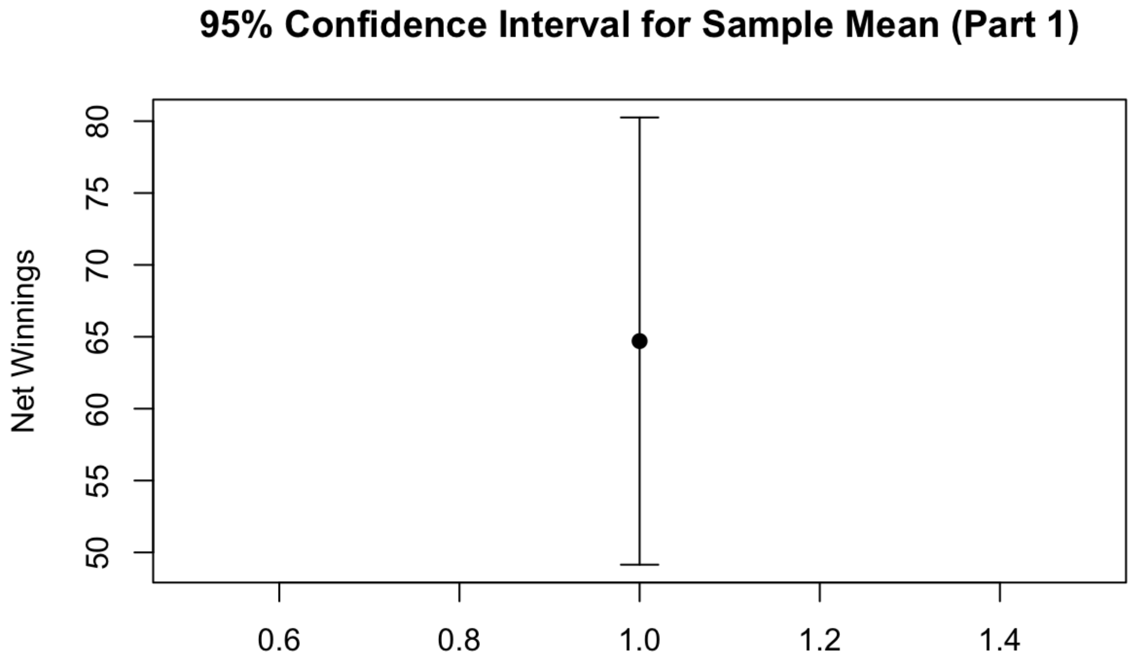
The betting strategy looks to be moderately favorable, with a predicted positive net win. However, the high standard deviation suggests a significant danger. While the method has promise, its success is contingent on individual risk tolerance.

Visualization:



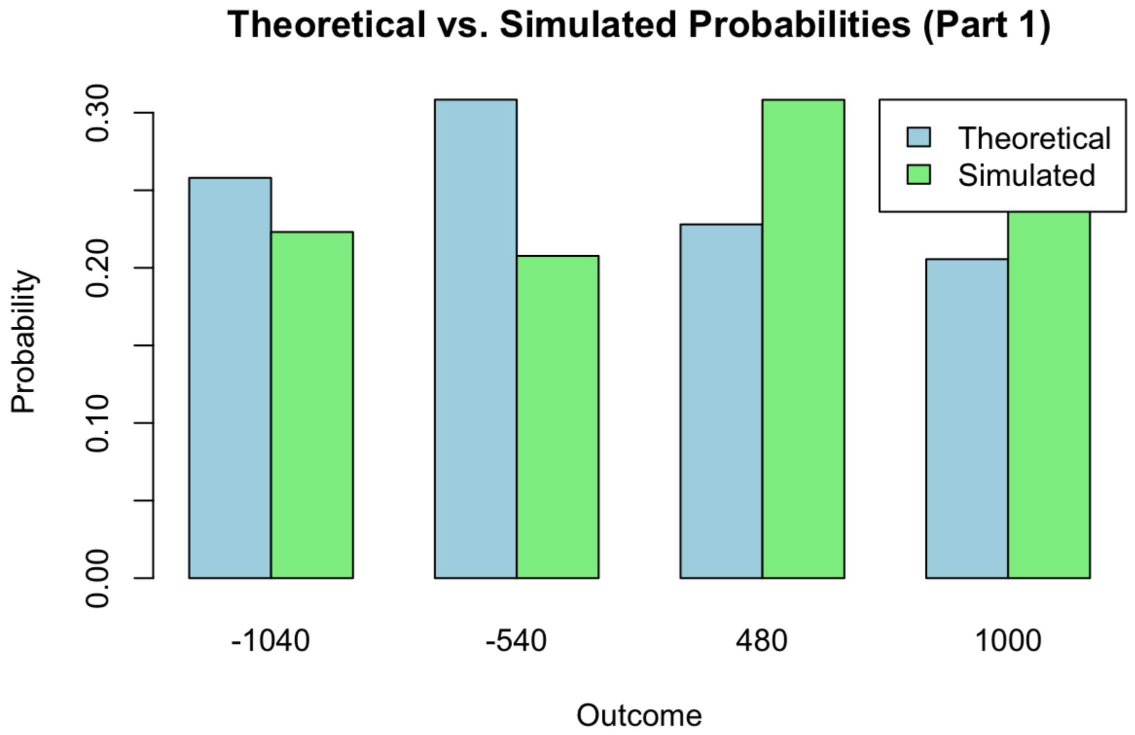
Interpretation:

The histogram shows that most results are concentrated around the specified numbers indicating net winnings, while the various peaks represent established winning and losing outcomes.



Interpretation:

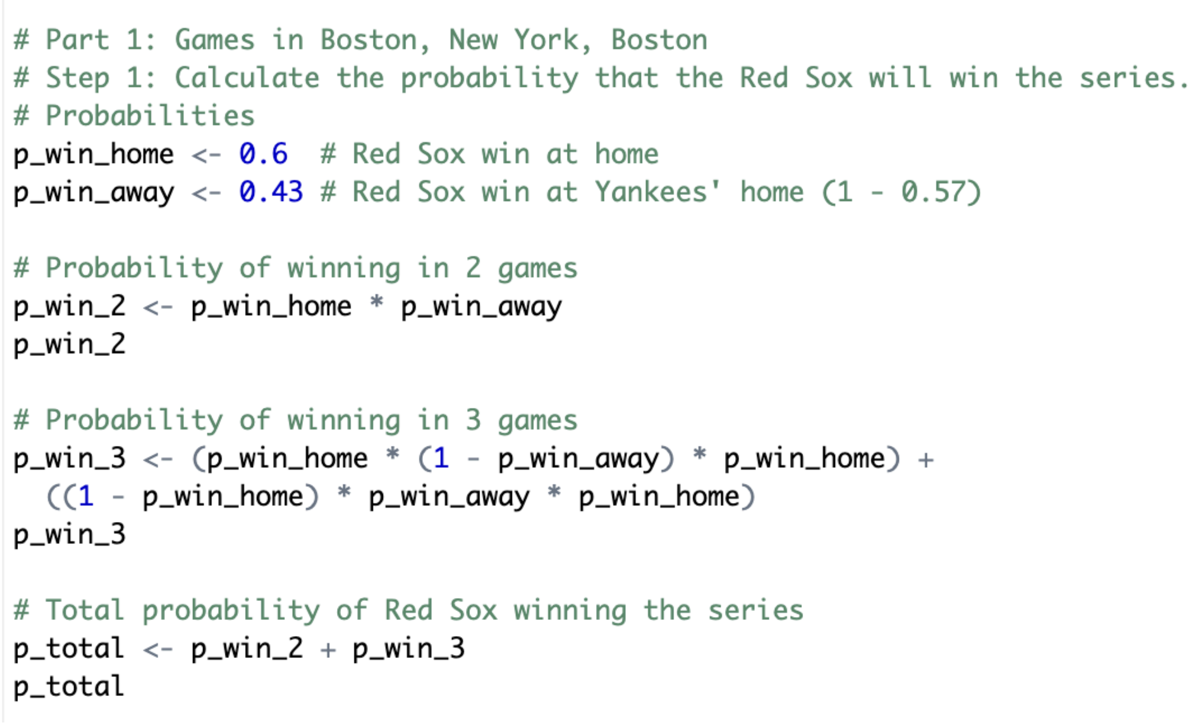
The confidence interval for the simulation lies between $49.15 and $80.26, which includes the theoretical expected value ($57.89). This statistic proves that both the theory and simulation are consistent with each other.



Interpretation:

On the surface, theoretical and simulated probabilities seem comparable, with their biggest discrepancy being in what one would expect to happen when playing three games, which was expected given the randomness in the results of simulation.

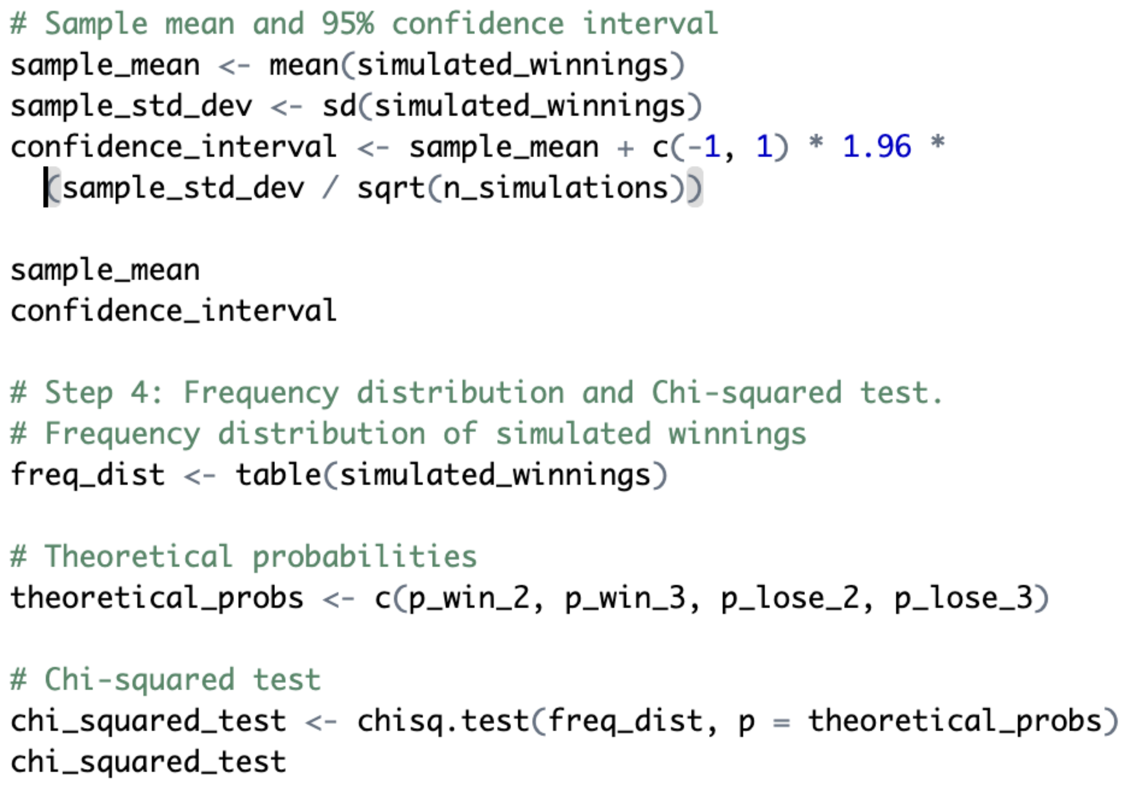
Code:



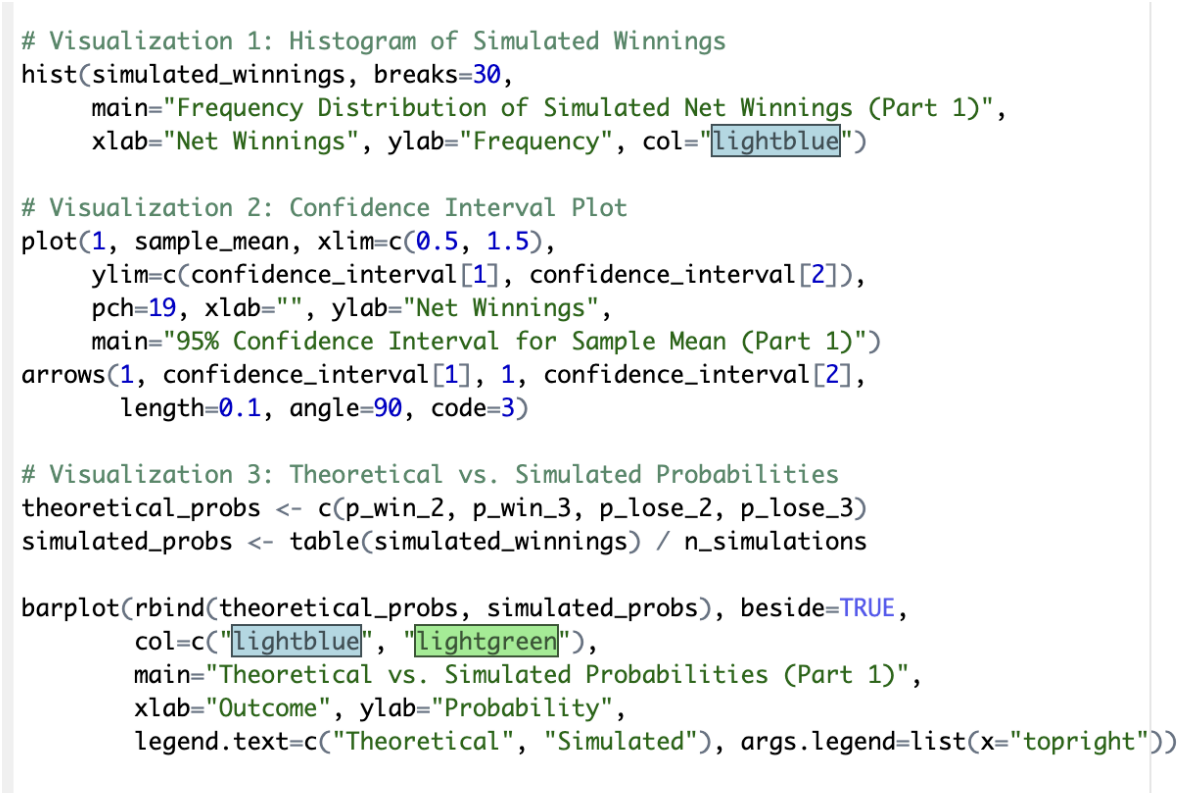
```
# Part 1: Games in Boston, New York, Boston
# Step 1: Calculate the probability that the Red Sox will win the series.
# Probabilities
p_win_home <- 0.6  # Red Sox win at home
p_win_away <- 0.43 # Red Sox win at Yankees' home (1 - 0.57)

# Probability of winning in 2 games
p_win_2 <- p_win_home * p_win_away
p_win_2

# Probability of winning in 3 games
p_win_3 <- (p_win_home * (1 - p_win_away) * p_win_home) +
  ((1 - p_win_home) * p_win_away * p_win_home)
p_win_3

# Total probability of Red Sox winning the series
p_total <- p_win_2 + p_win_3
p_total
```

```r
# Step 2: Construct the theoretical probability distribution for net winnings (X).
# Net winnings for each outcome
winnings_win_2 <- 1000
winnings_win_3 <- 480
winnings_lose_2 <- -1040
winnings_lose_3 <- -540

# Probabilities for each outcome
p_win_2
p_win_3

p_lose_2 <- (1 - p_win_home) * (1 - p_win_away)
p_lose_2

p_lose_3 <- (p_win_home * (1 - p_win_away) * (1 - p_win_home)) +
  ((1 - p_win_home) * p_win_away * (1 - p_win_home))
p_lose_3

# Expected value (mean) of X
expected_value <- (winnings_win_2 * p_win_2) +
  (winnings_win_3 * p_win_3) +
  (winnings_lose_2 * p_lose_2) +
  (winnings_lose_3 * p_lose_3)

# Variance and standard deviation of X
variance <- ((winnings_win_2 - expected_value)^2 * p_win_2) +
  ((winnings_win_3 - expected_value)^2 * p_win_3) +
  ((winnings_lose_2 - expected_value)^2 * p_lose_2) +
  ((winnings_lose_3 - expected_value)^2 * p_lose_3)
std_dev <- sqrt(variance)

expected_value
std_dev
```

```r
# Step 3: Simulate 10,000 series and compute a 95% confidence interval.
set.seed(123) # For reproducibility
n_simulations <- 10000

# Function to simulate a series
simulate_series <- function() {
  games <- c("Boston", "New York", "Boston")
  red_sox_wins <- 0
  yankees_wins <- 0
  net_winnings <- 0

  for (game in games) {
    if (game == "Boston") {
      red_sox_win <- runif(1) < p_win_home
    } else {
      red_sox_win <- runif(1) < p_win_away
    }

    if (red_sox_win) {
      red_sox_wins <- red_sox_wins + 1
      net_winnings <- net_winnings + 500
    } else {
      yankees_wins <- yankees_wins + 1
      net_winnings <- net_winnings - 520
    }

    if (red_sox_wins == 2 || yankees_wins == 2) {
      break
    }
  }

  return(net_winnings)
}

# Simulate 10,000 series
simulated_winnings <- replicate(n_simulations, simulate_series())
```



```
# Sample mean and 95% confidence interval
sample_mean <- mean(simulated_winnings)
sample_std_dev <- sd(simulated_winnings)
confidence_interval <- sample_mean + c(-1, 1) * 1.96 *
  (sample_std_dev / sqrt(n_simulations))

sample_mean
confidence_interval

# Step 4: Frequency distribution and Chi-squared test.
# Frequency distribution of simulated winnings
freq_dist <- table(simulated_winnings)

# Theoretical probabilities
theoretical_probs <- c(p_win_2, p_win_3, p_lose_2, p_lose_3)

# Chi-squared test
chi_squared_test <- chisq.test(freq_dist, p = theoretical_probs)
chi_squared_test
```



```
# Visualization 1: Histogram of Simulated Winnings
hist(simulated_winnings, breaks=30,
     main="Frequency Distribution of Simulated Net Winnings (Part 1)",
     xlab="Net Winnings", ylab="Frequency", col="lightblue")

# Visualization 2: Confidence Interval Plot
plot(1, sample_mean, xlim=c(0.5, 1.5),
     ylim=c(confidence_interval[1], confidence_interval[2]),
     pch=19, xlab="", ylab="Net Winnings",
     main="95% Confidence Interval for Sample Mean (Part 1)")
arrows(1, confidence_interval[1], 1, confidence_interval[2],
       length=0.1, angle=90, code=3)

# Visualization 3: Theoretical vs. Simulated Probabilities
theoretical_probs <- c(p_win_2, p_win_3, p_lose_2, p_lose_3)
simulated_probs <- table(simulated_winnings) / n_simulations

barplot(rbind(theoretical_probs, simulated_probs), beside=TRUE,
        col=c("lightblue", "lightgreen"),
        main="Theoretical vs. Simulated Probabilities (Part 1)",
        xlab="Outcome", ylab="Probability",
        legend.text=c("Theoretical", "Simulated"), args.legend=list(x="topright"))
```

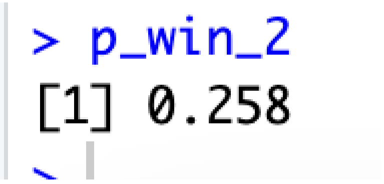
Part 2: Games in New York, Boston, New York

1. **Step 1: Calculate the probability that the Red Sox will win the series.**

I compute the total probability of the Red Sox winning the series by evaluating two scenarios:

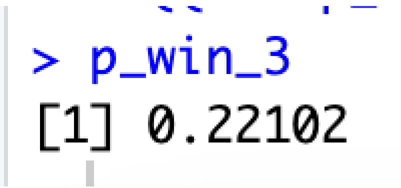
- Winning in 2 games: P(Win in 2 games) = P(Win away) × P(Win at home)

= 0.43 × 0.6 = 0.258



```
> p_win_2
[1] 0.258
```

- Winning in 3 games: P(Win in 3 games) = (0.43 x 0.4 x 0.43) + (0.57 x 0.6 x 0.43) = 0.22102.

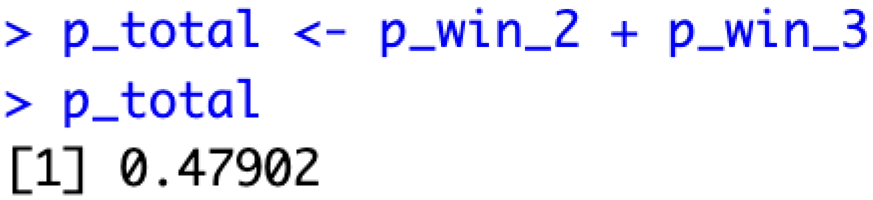


```
> p_win_3
[1] 0.22102
```

Total probability:

P(Total) = P(Win in 2 games) + P(Win in 3 games) = 0.258 + 0.22102 = 0.47902.

Output:



```
> p_total <- p_win_2 + p_win_3
> p_total
[1] 0.47902
```

Interpretation:

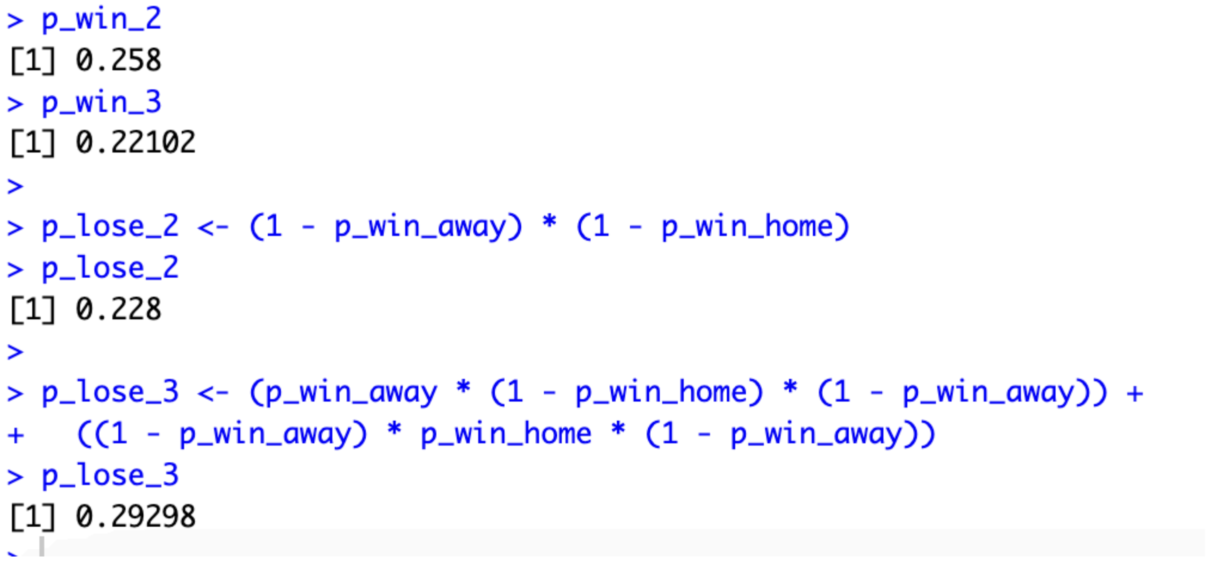
The Red Sox's probability of winning the series decreases to 47.90% when the first and last games are played away.

2. **Step 2: Construct the theoretical probability distribution for net winnings (X).**

Similar to Part 1, I formulate the theoretical distribution of net wins X for each potential outcome:

Outcome	Net Winnings ($)	Probability
Win in 2 games	+1000	0.258
Win in 3 games	+480	0.22102
Lose in 2 games	-1040	0.228
Lose in 3 games	-540	0.29209

Output:

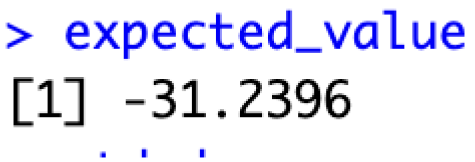


```
> p_win_2
[1] 0.258
> p_win_3
[1] 0.22102
>
> p_lose_2 <- (1 - p_win_away) * (1 - p_win_home)
> p_lose_2
[1] 0.228
>
> p_lose_3 <- (p_win_away * (1 - p_win_home) * (1 - p_win_away)) +
+   ((1 - p_win_away) * p_win_home * (1 - p_win_away))
> p_lose_3
[1] 0.29298
```

Expected Value (Mean of X):

E(X) = (1000 × 0.258) + (480 × 0.22102) + (−1040 × 0.228) + (−540 × 0.29298) = -31.24

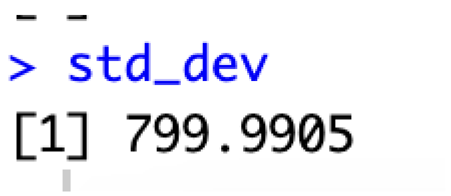
Output:



```
> expected_value
[1] -31.2396
```

Standard Deviation: 799.99

Output:



```
> std_dev
[1] 799.9905
```

Interpretation:

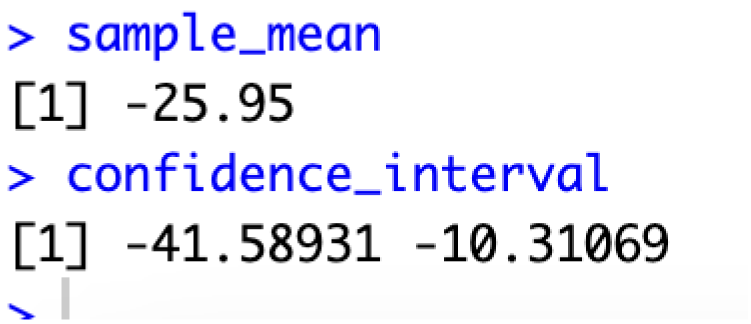
Compared to Part 1, when the predicted net profits were positive, the expected value is now negative ($-31.24), showing that this betting approach is less favorable when the series begins and ends in New York.

3. **Step 3: Simulate 10,000 series and compute a 95% confidence interval.**

Similar to Part 1, I have sample mean and 95% confidence interval as below:

- Sample Mean: $-25.95
- 95% Confidence Interval: [$-41.59, $-10.31]

Output:



```
> sample_mean
[1] -25.95
> confidence_interval
[1] -41.58931 -10.31069
>
```

Interpretation:

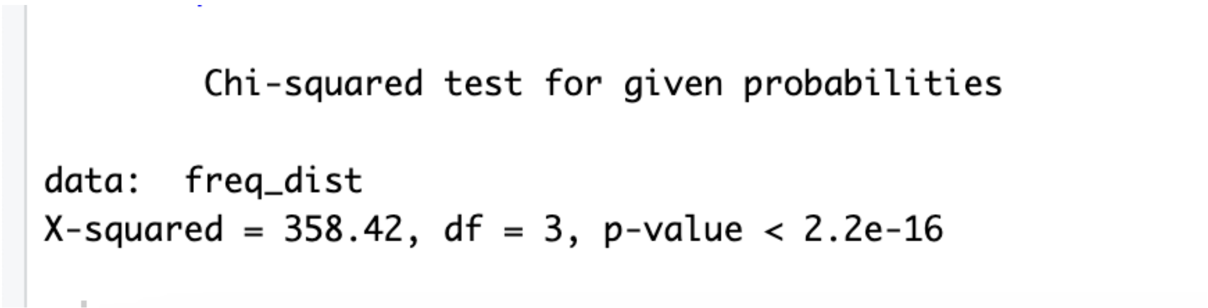
As we can see, the confidence interval is totally negative, which indicates that the betting strategy would be unfavorable for this game sequence..

4. **Step 4: Frequency distribution and Chi-squared test.**

Similar to Part 1, I have the result of Chi-squared test and p-value as below:

- Chi-squared Statistic: 358.42
- p-value: < 2.2e-16

Output:



```
        Chi-squared test for given probabilities

data:  freq_dist
X-squared = 358.42, df = 3, p-value < 2.2e-16
```

Interpretation:

The high p-value shows that the simulated results deviate significantly from the theoretical distribution due to randomness or model assumptions.

5. **Step 5: Determine if my betting strategy is favorable.**

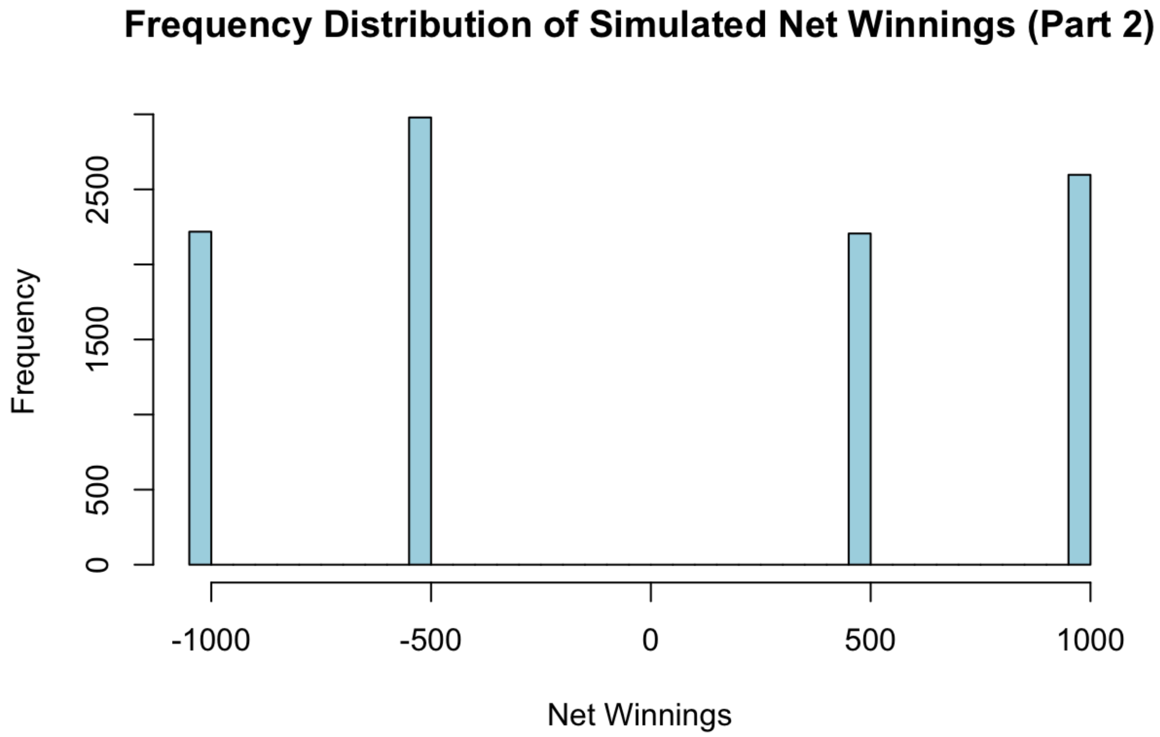
My observation through steps 2 and 3:

- Step 2: The expected net earnings are negative ($-31.24), making this a poor betting strategy.
- Step 3: The confidence interval is totally negative, indicating that the player is likely to lose money in this situation.

Final Interpretation:

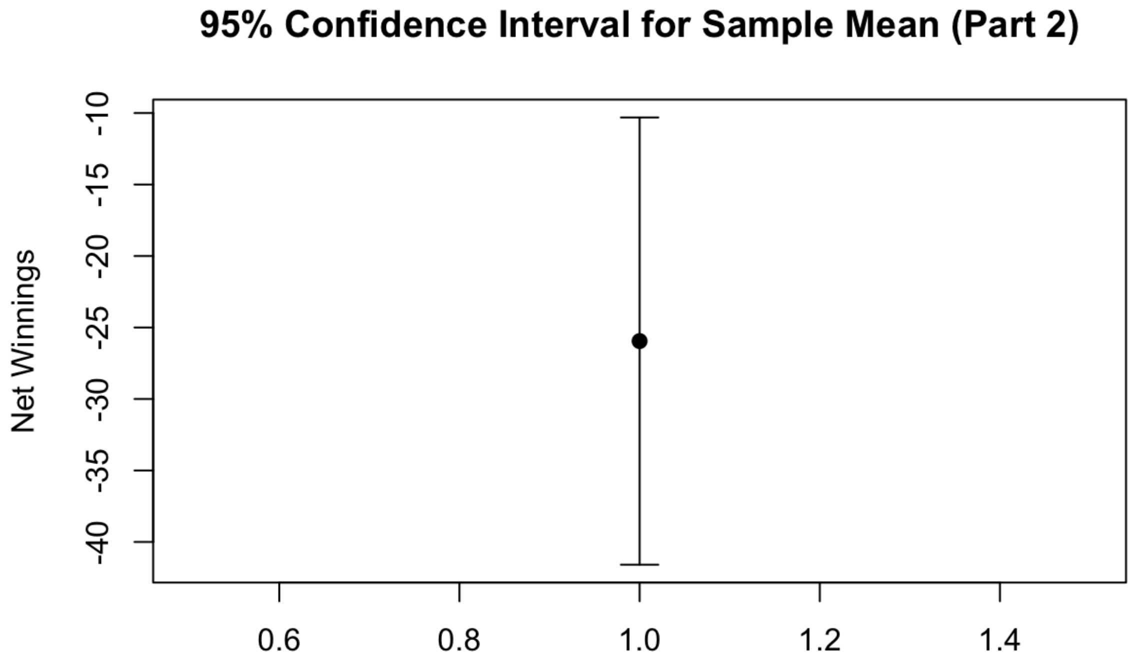
Changing the game order to New York, Boston, New York has a negative influence on the betting strategy. Unlike Part 1, when the expected winnings were positive, this sequence has a larger probability of losing and an expected net loss.

Visualization:



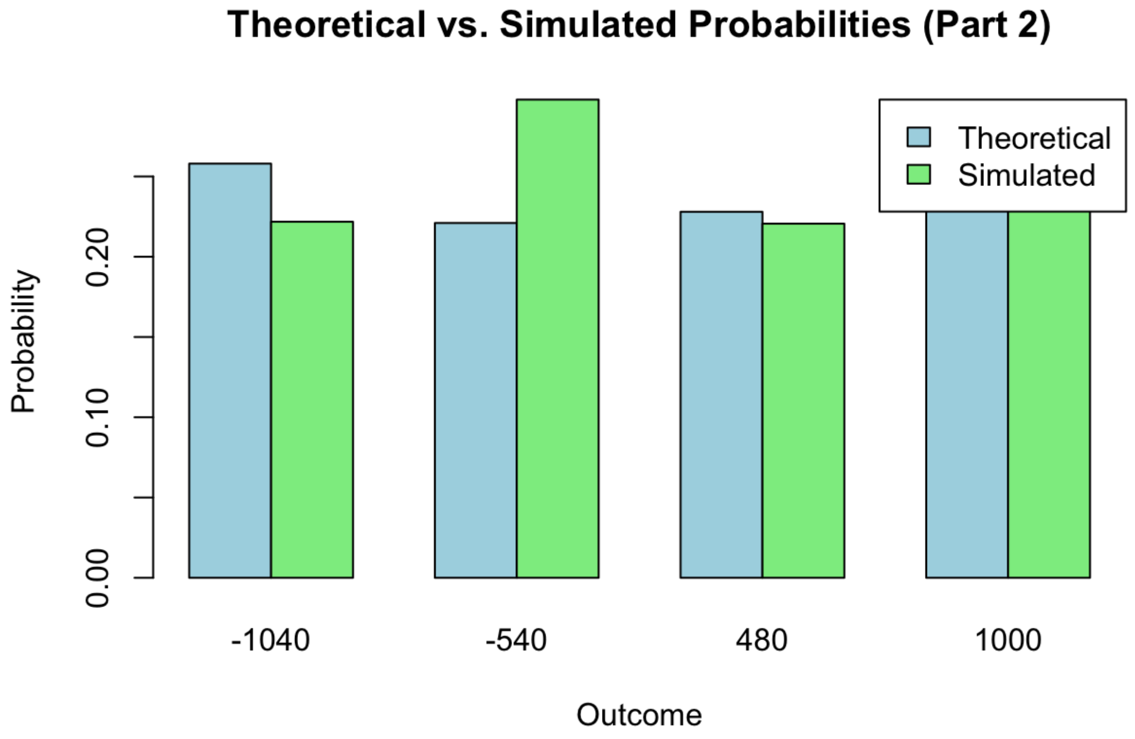
Interpretation:

The histogram reveals distinct peaks at the potential net wins, demonstrating that the series results are still grouped around four primary scenarios.



Interpretation:

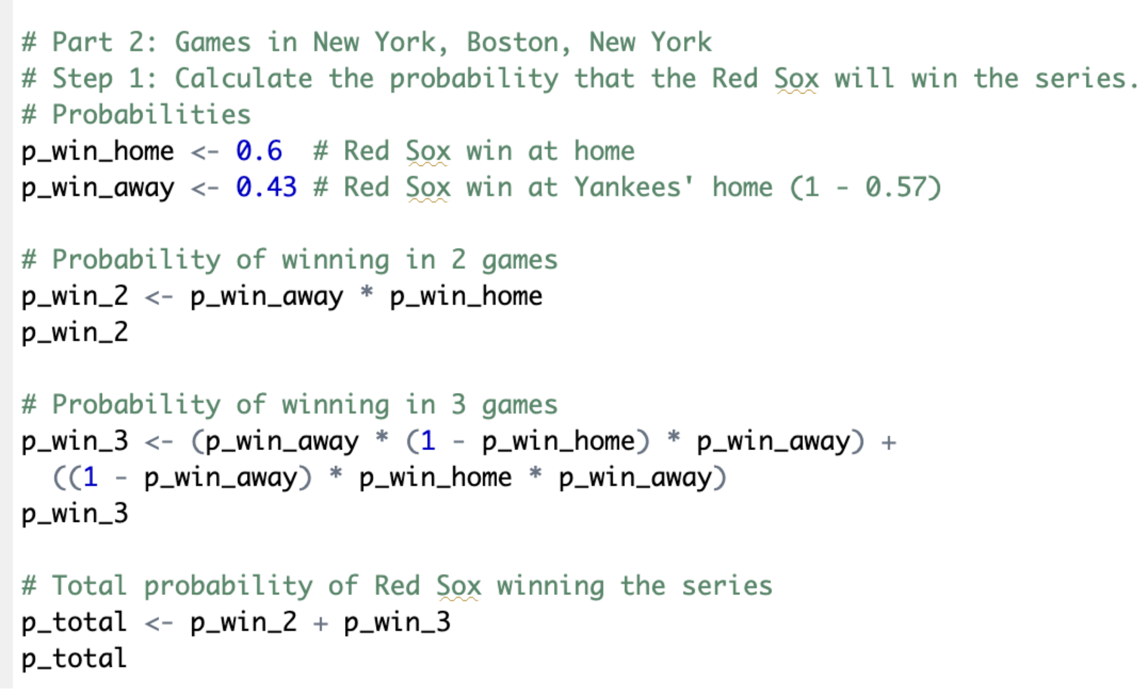
The confidence interval is totally below zero, indicating that this betting strategy will likely result in financial losses.



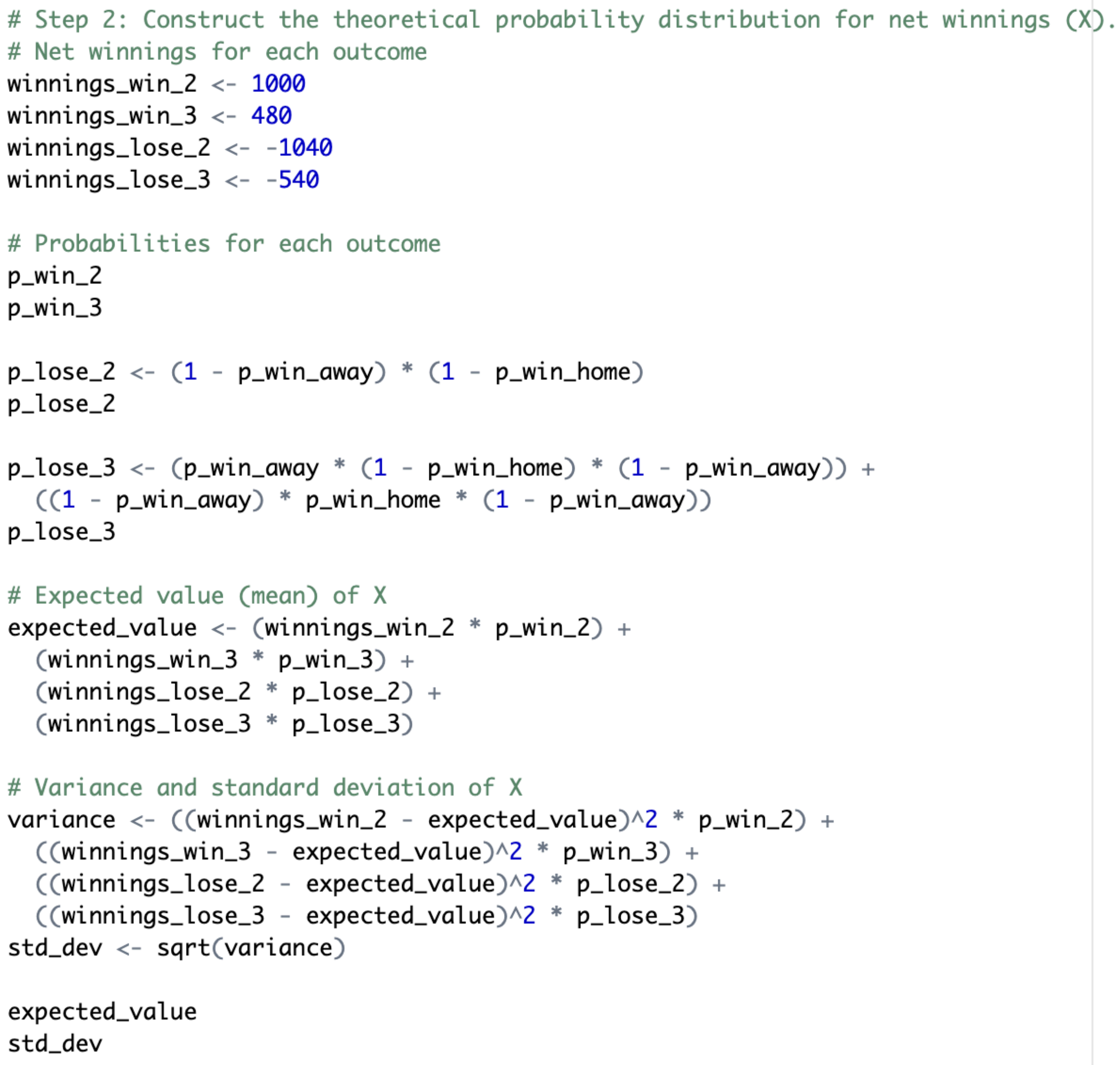
Interpretation:

The simulated probabilities are quite close to the theoretical values, with only some deviations due to simulation randomness.

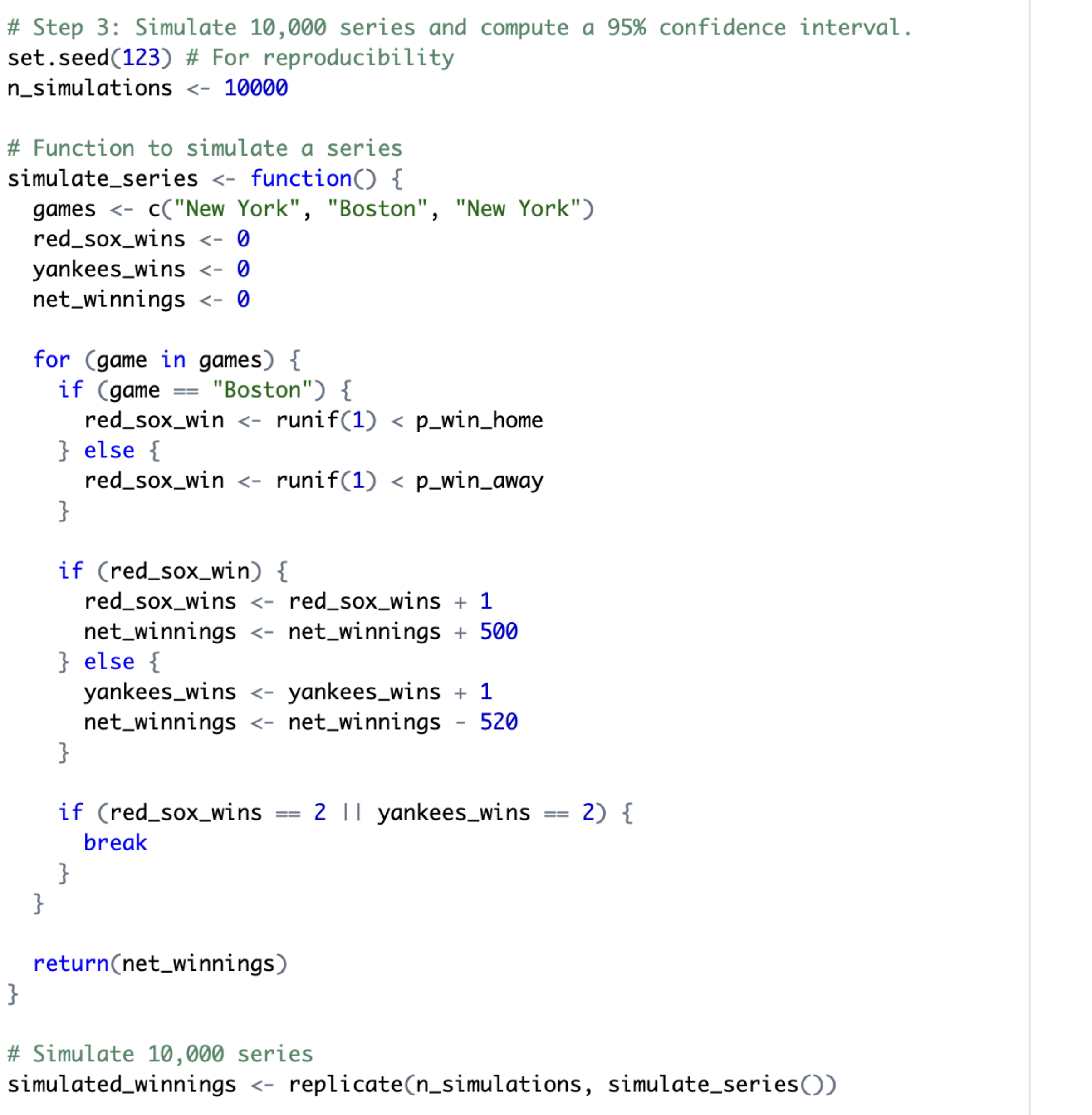
Code:



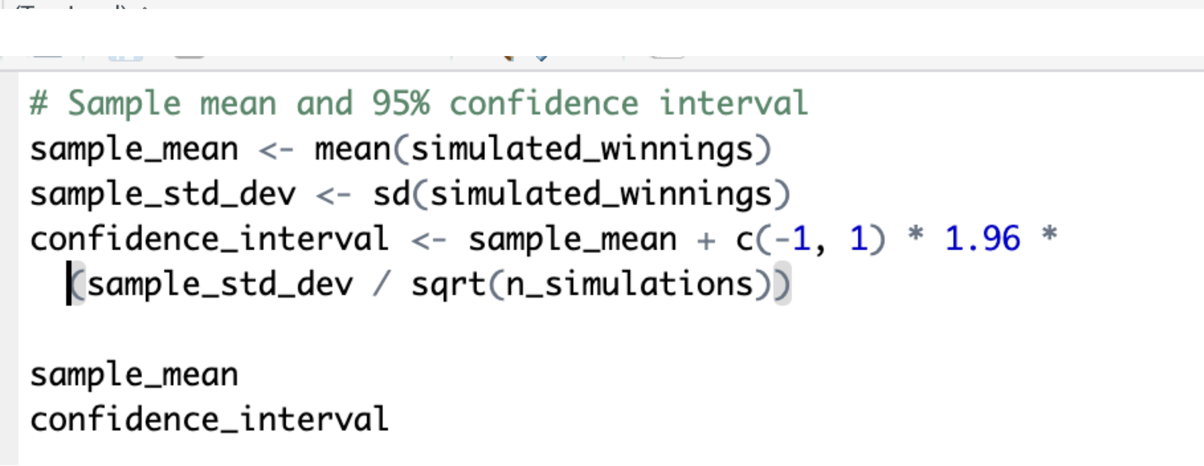
```
# Part 2: Games in New York, Boston, New York
# Step 1: Calculate the probability that the Red Sox will win the series.
# Probabilities
p_win_home <- 0.6  # Red Sox win at home
p_win_away <- 0.43 # Red Sox win at Yankees' home (1 - 0.57)

# Probability of winning in 2 games
p_win_2 <- p_win_away * p_win_home
p_win_2

# Probability of winning in 3 games
p_win_3 <- (p_win_away * (1 - p_win_home) * p_win_away) +
  ((1 - p_win_away) * p_win_home * p_win_away)
p_win_3

# Total probability of Red Sox winning the series
p_total <- p_win_2 + p_win_3
p_total
```



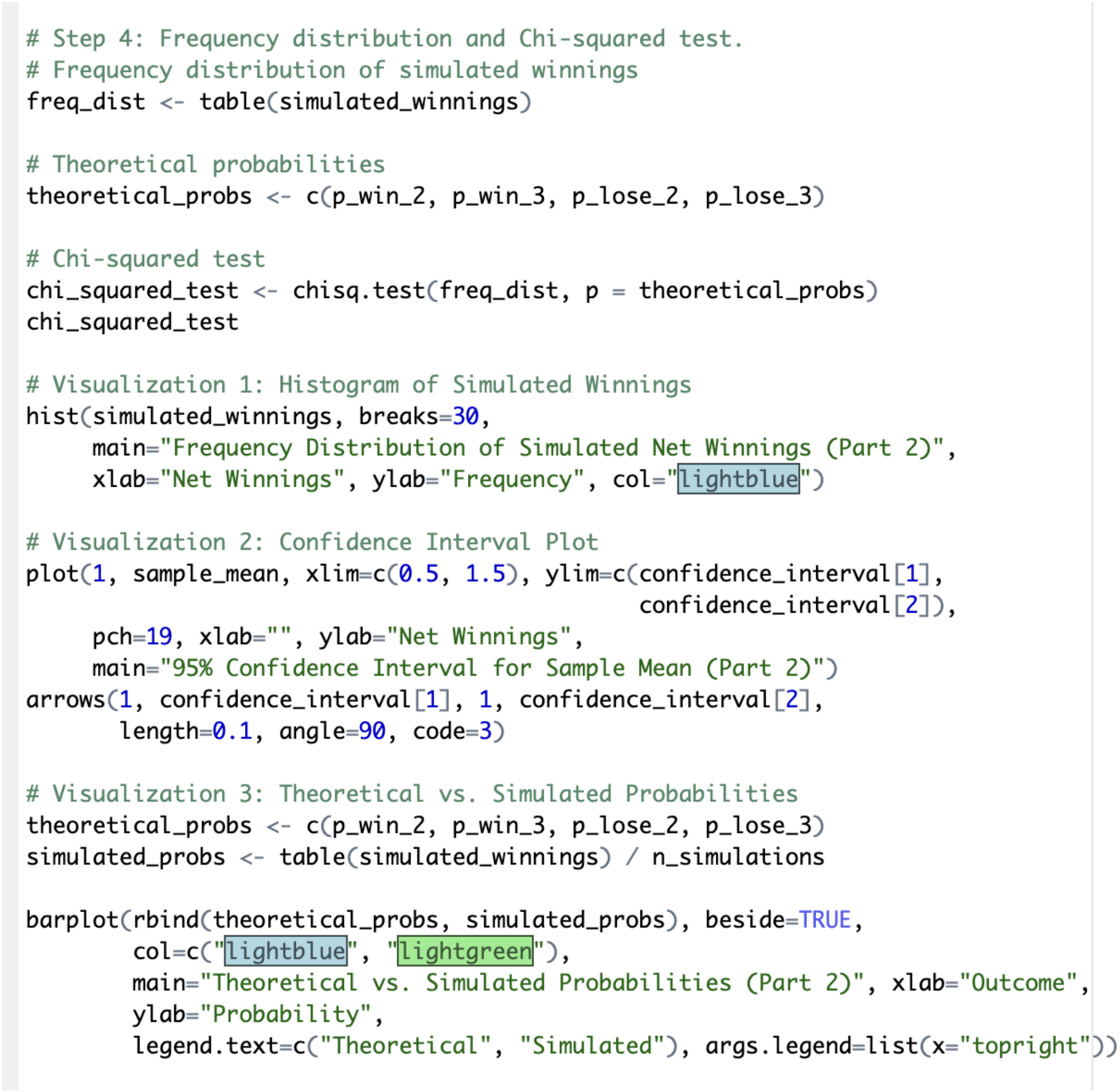
```r
# Step 2: Construct the theoretical probability distribution for net winnings (X).
# Net winnings for each outcome
winnings_win_2 <- 1000
winnings_win_3 <- 480
winnings_lose_2 <- -1040
winnings_lose_3 <- -540

# Probabilities for each outcome
p_win_2
p_win_3

p_lose_2 <- (1 - p_win_away) * (1 - p_win_home)
p_lose_2

p_lose_3 <- (p_win_away * (1 - p_win_home) * (1 - p_win_away)) +
  ((1 - p_win_away) * p_win_home * (1 - p_win_away))
p_lose_3

# Expected value (mean) of X
expected_value <- (winnings_win_2 * p_win_2) +
  (winnings_win_3 * p_win_3) +
  (winnings_lose_2 * p_lose_2) +
  (winnings_lose_3 * p_lose_3)

# Variance and standard deviation of X
variance <- ((winnings_win_2 - expected_value)^2 * p_win_2) +
  ((winnings_win_3 - expected_value)^2 * p_win_3) +
  ((winnings_lose_2 - expected_value)^2 * p_lose_2) +
  ((winnings_lose_3 - expected_value)^2 * p_lose_3)
std_dev <- sqrt(variance)

expected_value
std_dev
```



```
# Step 3: Simulate 10,000 series and compute a 95% confidence interval.
set.seed(123) # For reproducibility
n_simulations <- 10000

# Function to simulate a series
simulate_series <- function() {
  games <- c("New York", "Boston", "New York")
  red_sox_wins <- 0
  yankees_wins <- 0
  net_winnings <- 0

  for (game in games) {
    if (game == "Boston") {
      red_sox_win <- runif(1) < p_win_home
    } else {
      red_sox_win <- runif(1) < p_win_away
    }

    if (red_sox_win) {
      red_sox_wins <- red_sox_wins + 1
      net_winnings <- net_winnings + 500
    } else {
      yankees_wins <- yankees_wins + 1
      net_winnings <- net_winnings - 520
    }

    if (red_sox_wins == 2 || yankees_wins == 2) {
      break
    }
  }

  return(net_winnings)
}

# Simulate 10,000 series
simulated_winnings <- replicate(n_simulations, simulate_series())
```



```
# Sample mean and 95% confidence interval
sample_mean <- mean(simulated_winnings)
sample_std_dev <- sd(simulated_winnings)
confidence_interval <- sample_mean + c(-1, 1) * 1.96 *
  (sample_std_dev / sqrt(n_simulations))

sample_mean
confidence_interval
```



```
# Step 4: Frequency distribution and Chi-squared test.
# Frequency distribution of simulated winnings
freq_dist <- table(simulated_winnings)

# Theoretical probabilities
theoretical_probs <- c(p_win_2, p_win_3, p_lose_2, p_lose_3)

# Chi-squared test
chi_squared_test <- chisq.test(freq_dist, p = theoretical_probs)
chi_squared_test

# Visualization 1: Histogram of Simulated Winnings
hist(simulated_winnings, breaks=30,
     main="Frequency Distribution of Simulated Net Winnings (Part 2)",
     xlab="Net Winnings", ylab="Frequency", col="lightblue")

# Visualization 2: Confidence Interval Plot
plot(1, sample_mean, xlim=c(0.5, 1.5), ylim=c(confidence_interval[1],
                                              confidence_interval[2]),
     pch=19, xlab="", ylab="Net Winnings",
     main="95% Confidence Interval for Sample Mean (Part 2)")
arrows(1, confidence_interval[1], 1, confidence_interval[2],
       length=0.1, angle=90, code=3)

# Visualization 3: Theoretical vs. Simulated Probabilities
theoretical_probs <- c(p_win_2, p_win_3, p_lose_2, p_lose_3)
simulated_probs <- table(simulated_winnings) / n_simulations

barplot(rbind(theoretical_probs, simulated_probs), beside=TRUE,
        col=c("lightblue", "lightgreen"),
        main="Theoretical vs. Simulated Probabilities (Part 2)", xlab="Outcome",
        ylab="Probability",
        legend.text=c("Theoretical", "Simulated"), args.legend=list(x="topright"))
```

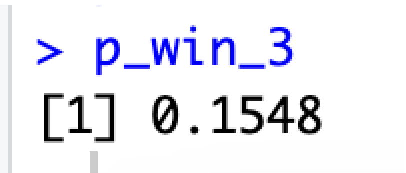
Part 3: Best-of-Five Series

1. **Step 1: Calculate the probability that the Red Sox will win the series.**

Similar to Part 1, 2, I have a probability of winning 3, 4, and 5 games as below:

- Winning in 3 games: P(Win in 3 games) = P(Win at home) × P(Win away) × P(Win at home) = 0.6 × 0.43 × 0.6 = 0.1548

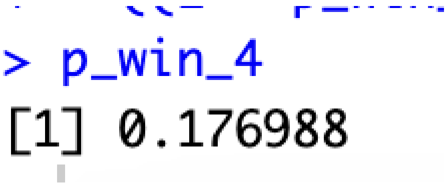
Output:



```
> p_win_3
[1] 0.1548
```

- Winning in 4 games: P(Win in 4 games) = (0.6 × 0.43 × 0.4 × 0.43) + (0.6 × 0.57 × 0.6 × 0.43) + (0.4 × 0.43 × 0.6 × 0.43) = 0.177

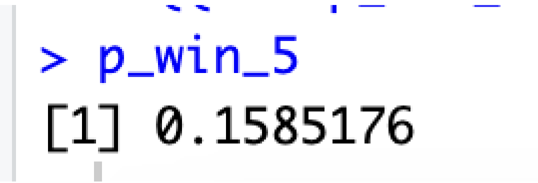
Output:



```
> p_win_4
[1] 0.176988
```

- Winning in 5 games: P(Win in 5 games) = (0.6 × 0.43 × 0.4 × 0.57 × 0.6) + (0.6 × 0.57 × 0.6 × 0.57 × 0.6) + (0.4 × 0.43 × 0.6 × 0.57 × 0.6) + (0.4 × 0.43 × 0.4 × 0.43 × 0.6) = 0.1585

Output:

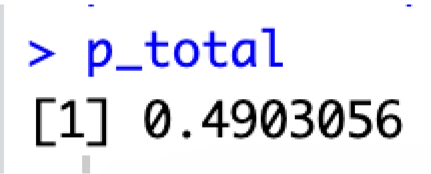


```
> p_win_5
[1] 0.1585176
```

Total probability:

P(Total) = P(Win in 3 games) + P(Win in 4 games) + P(Win in 5 games) = 0.1548 + 0.177 + 0.1585 = 0.4903

Output:



```
> p_total
[1] 0.4903056
```

Interpretation:

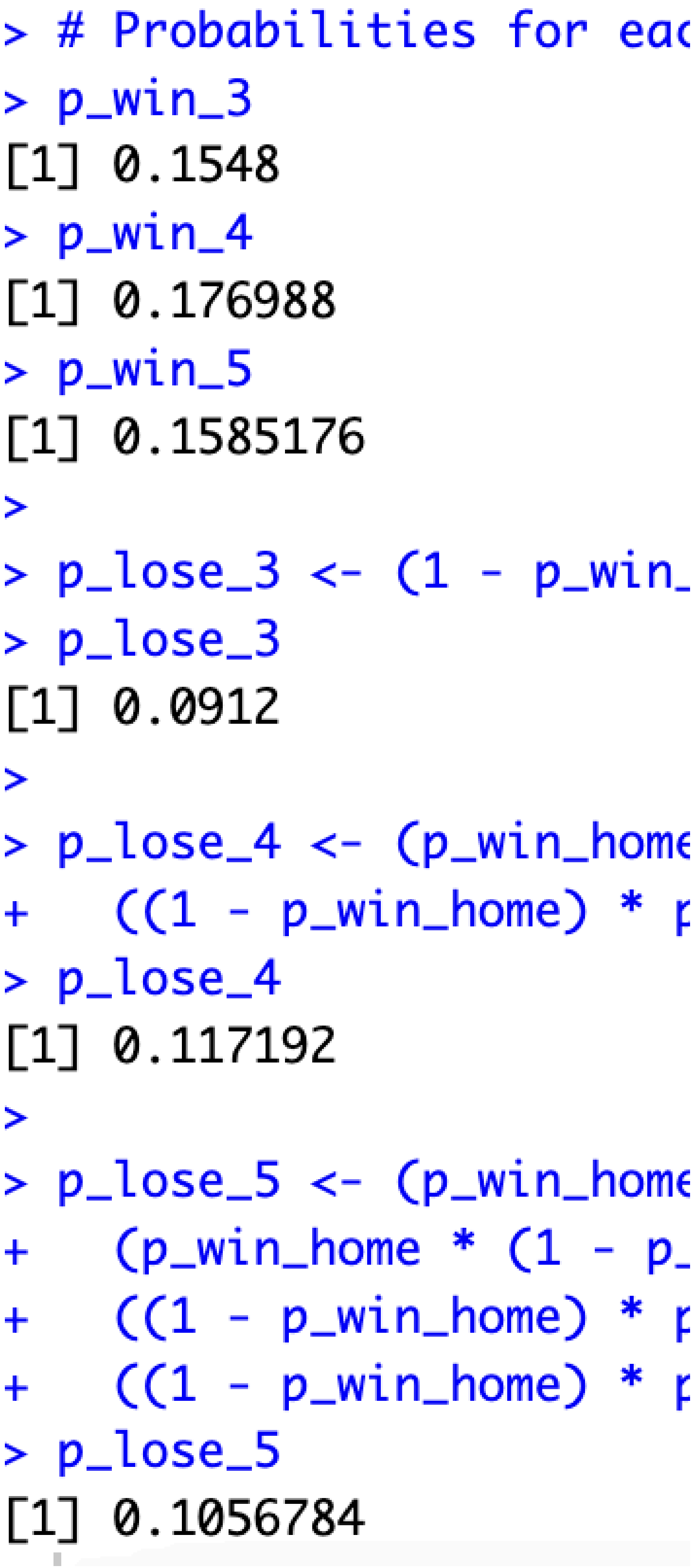
The Red Sox have a 49.03% chance of winning the series; therefore, the Yankees have a slightly better chance of winning in this scenario.

2. **Step 2: Construct the theoretical probability distribution for net winnings (X).**

Similar to Part 1, and 2, I formulate the theoretical distribution of net wins X for each potential outcome:

Outcome	Net Winnings ($)	Probability
Win in 3 games	+1500	0.1548
Win in 4 games	+980	0.177
Win in 4 games	+460	0.1585
Lose in 3 games	-1560	0.0912
Lose in 2 games	-1060	0.1172
Lose in 3 games	-560	0.1057

Output:

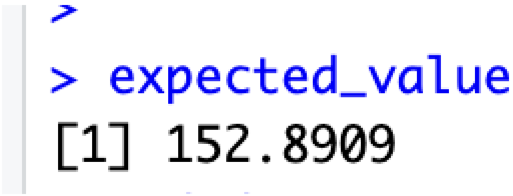


```
> # Probabilities for eac
> p_win_3
[1] 0.1548
> p_win_4
[1] 0.176988
> p_win_5
[1] 0.1585176
>
> p_lose_3 <- (1 - p_win_
> p_lose_3
[1] 0.0912
>
> p_lose_4 <- (p_win_home
+   ((1 - p_win_home) * p
> p_lose_4
[1] 0.117192
>
> p_lose_5 <- (p_win_home
+   (p_win_home * (1 - p_
+   ((1 - p_win_home) * p
+   ((1 - p_win_home) * p
> p_lose_5
[1] 0.1056784
```

Expected Value (Mean of X):

E(X) = (1500 × 0.1548) + (980 × 0.176988) + (460 × 0.1585) + (−1560 × 0.0912) + (−1060 × 0.1172) + (−560 × 0.1057) = 152.89

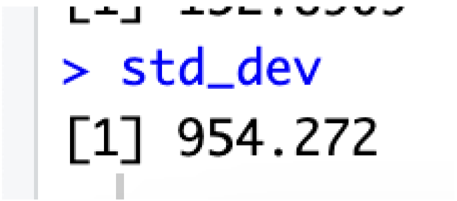
Output:



```
>
> expected_value
[1] 152.8909
```

Standard Deviation: 954.27

Output:



```
> std_dev
[1] 954.272
```

Interpretation:

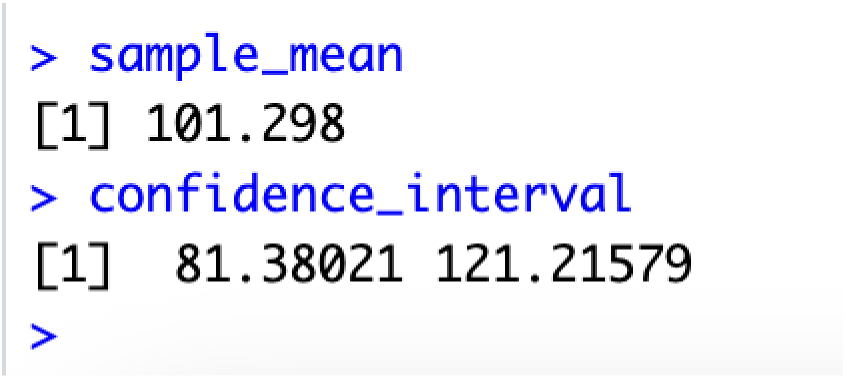
- The expected net winnings of 152.89 show that the betting strategy is more beneficial in a best-of-five series than in Parts 1 and 2.
- However, the large standard deviation ($954.27) indicates that results are very changeable, making this a dangerous investment.

3. **Step 3: Simulate 10,000 series and compute a 95% confidence interval.**

Similar to Part 1 and 2, I have sample mean and 95% confidence interval as below:

- Sample Mean: $101.3
- 95% Confidence Interval: [$81.38, $121.22]

Output:



```
> sample_mean
[1] 101.298
> confidence_interval
[1]  81.38021 121.21579
>
```

Interpretation:

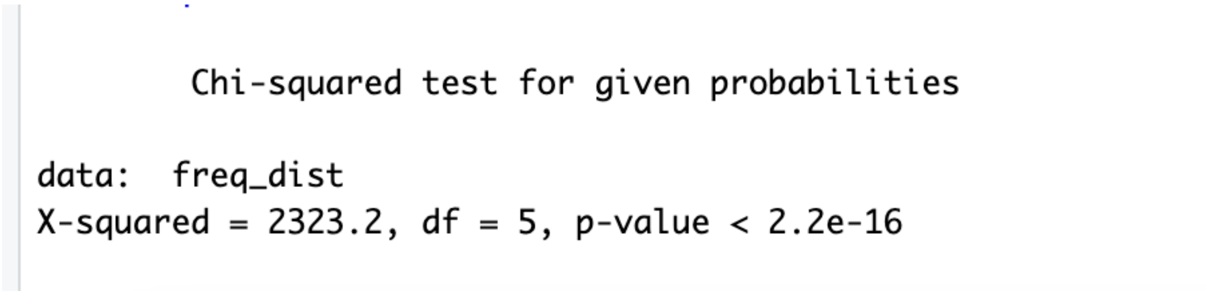
- The simulated mean of $101.30 is lower than the theoretical expectation ($152.89), indicating that random variations have a slight influence on the result.
- The confidence interval stays positive, indicating that the method is typically effective.

4. **Step 4: Frequency distribution and Chi-squared test.**

Similar to Part 1 and 2, I have the result of Chi-squared test and p-value as below:

- Chi-squared Statistic: 2323.2
- p-value: < 2.2e-16

Output:



```
	Chi-squared test for given probabilities

data:  freq_dist
X-squared = 2323.2, df = 5, p-value < 2.2e-16
```

Interpretation:

- The high Chi-squared test shows that the simulated results are not the same as the theoretical distribution because of the randomness of the samples.
- However, since the expected and simulated means are similar, the strategy stays statistically valid.

5. **Step 5: Determine if my betting strategy is favorable.**

My observation through steps 2 and 3:

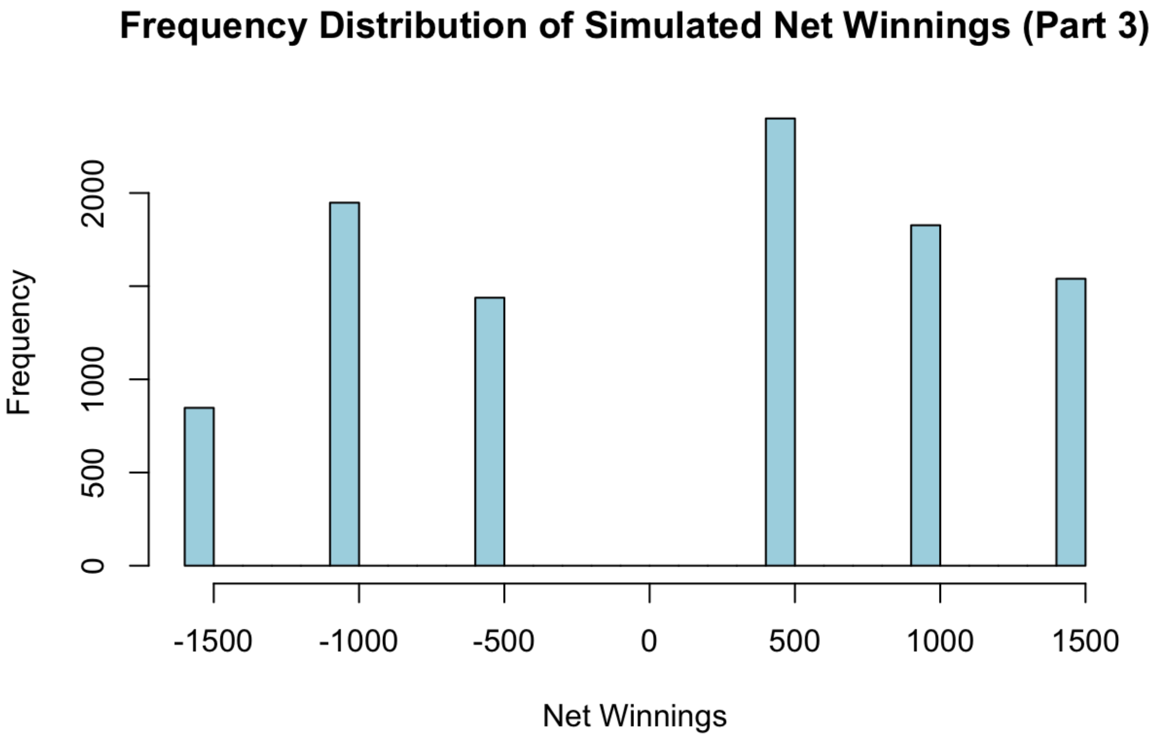
- The expected net winnings reached $152.89, making this betting strategy more profitable than the best-of-three format.
- The simulated outcomes closely match theory, and the positive confidence interval indicates a high possibility of profit.

- The risk remains high because of the high standard deviation ($954.27), which indicates that substantial losses are possible.

Final Interpretation:

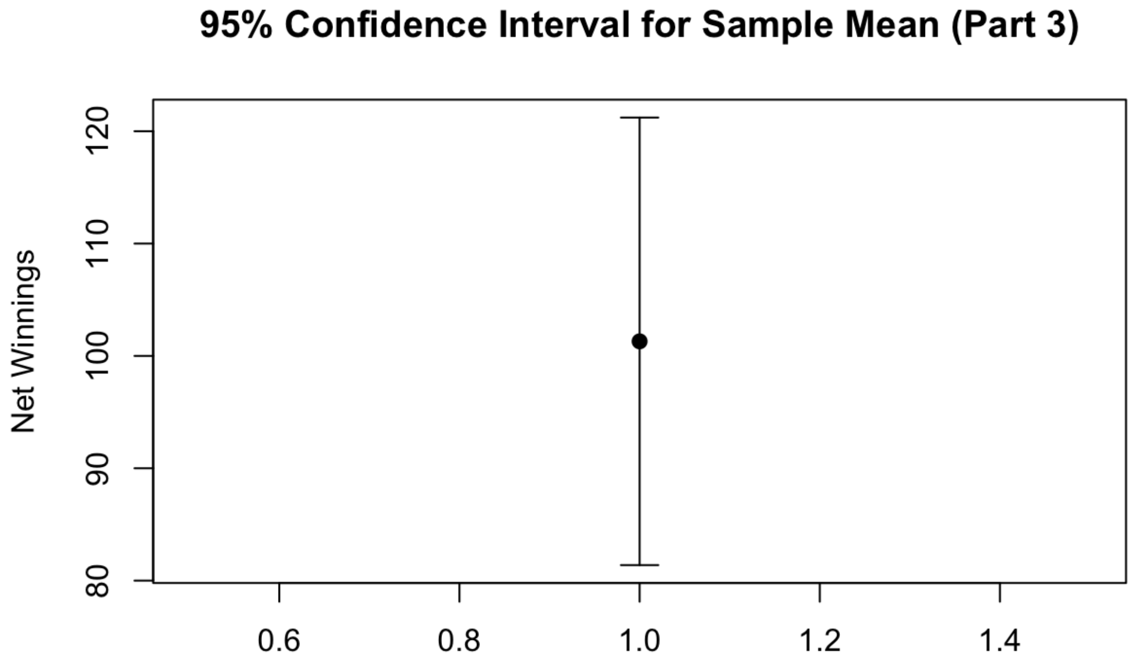
- The best-of-five series format enhances betting strategy by raising predicted wins.
- However, due to the significant variability, this approach remains dangerous and should be undertaken with caution.

Visualization:



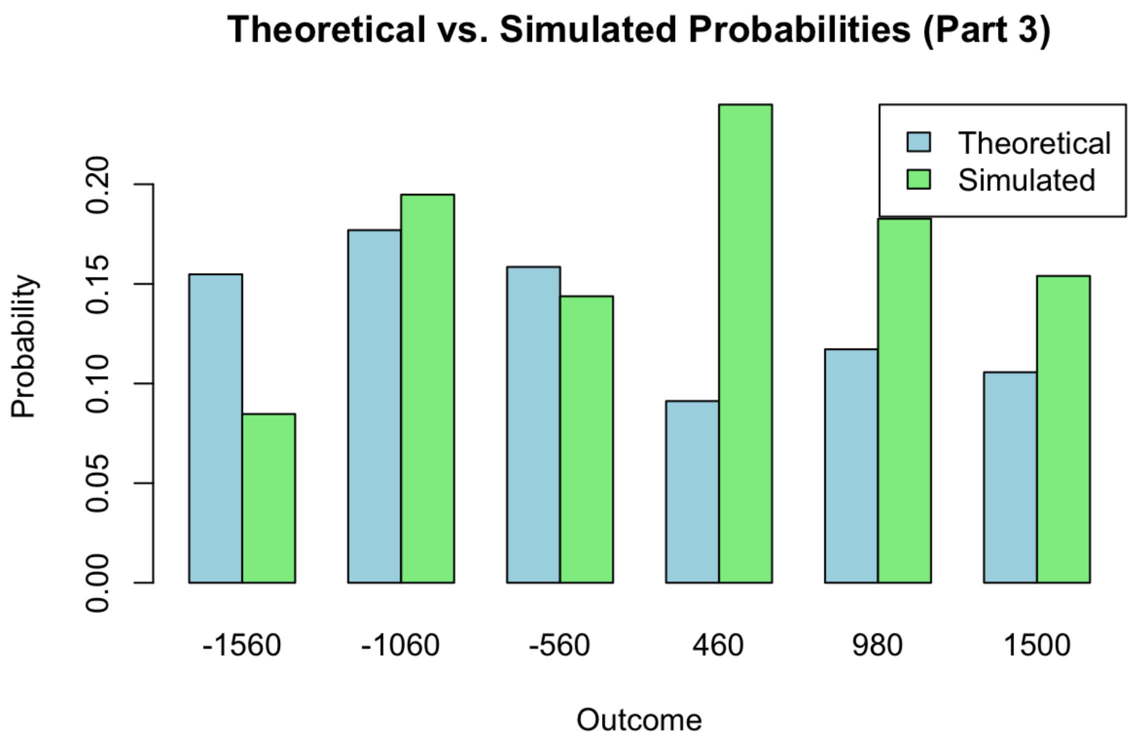
Interpretation:

- The histogram exhibits unique peaks for each possible net winnings level.
- The distribution demonstrates that outcomes are discrete, as predicted given the betting conditions.



Interpretation:

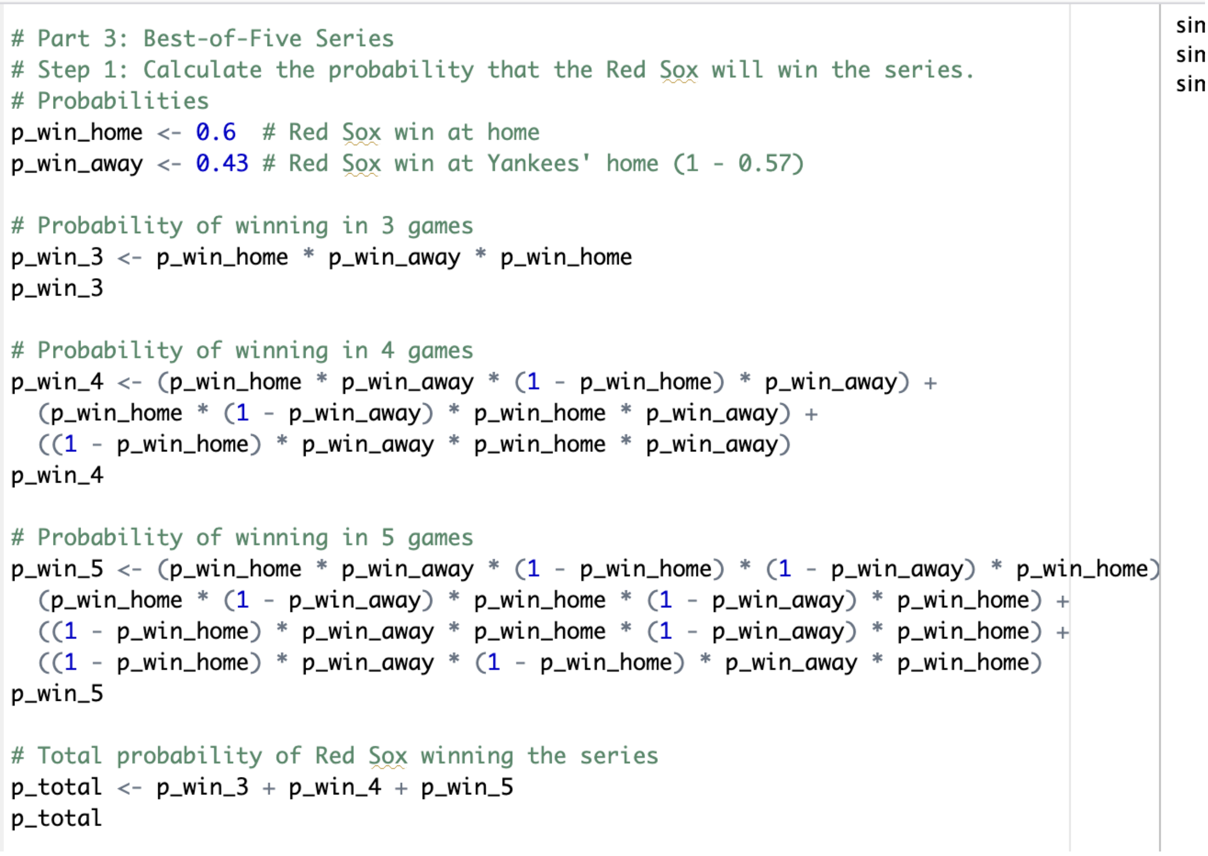
The confidence interval is fully above zero, indicating that the approach is likely to be favorable across a large number of bets.



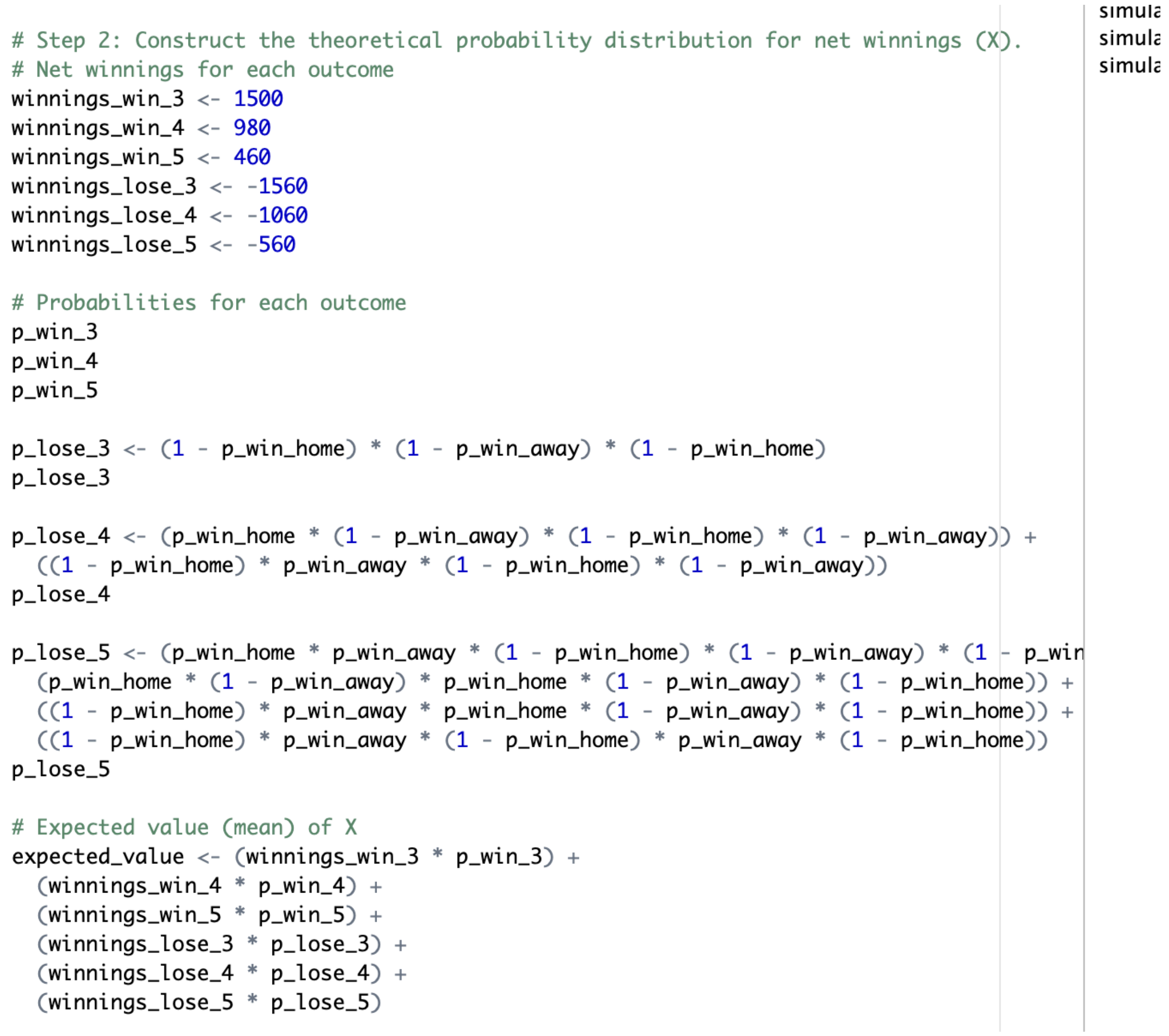
Interpretation:

- The simulated probabilities closely match the theoretical predictions, supporting the probability model.
- Small differences arise as a result of random variation in simulations.

Code:

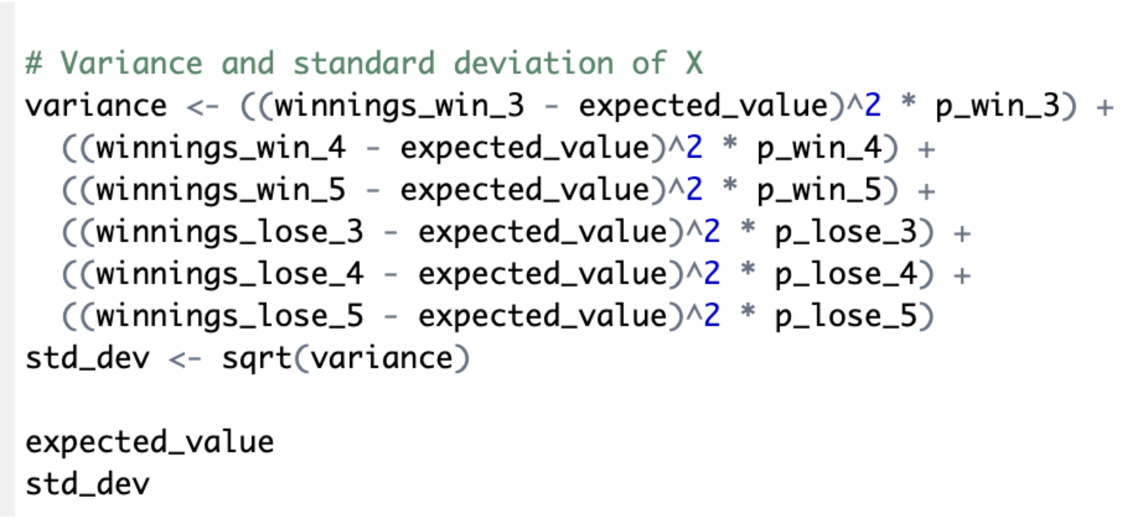


```
# Part 3: Best-of-Five Series
# Step 1: Calculate the probability that the Red Sox will win the series.
# Probabilities
p_win_home <- 0.6  # Red Sox win at home
p_win_away <- 0.43 # Red Sox win at Yankees' home (1 - 0.57)

# Probability of winning in 3 games
p_win_3 <- p_win_home * p_win_away * p_win_home
p_win_3

# Probability of winning in 4 games
p_win_4 <- (p_win_home * p_win_away * (1 - p_win_home) * p_win_away) +
  (p_win_home * (1 - p_win_away) * p_win_home * p_win_away) +
  ((1 - p_win_home) * p_win_away * p_win_home * p_win_away)
p_win_4

# Probability of winning in 5 games
p_win_5 <- (p_win_home * p_win_away * (1 - p_win_home) * (1 - p_win_away) * p_win_home)
  (p_win_home * (1 - p_win_away) * p_win_home * (1 - p_win_away) * p_win_home) +
  ((1 - p_win_home) * p_win_away * p_win_home * (1 - p_win_away) * p_win_home) +
  ((1 - p_win_home) * p_win_away * (1 - p_win_home) * p_win_away * p_win_home)
p_win_5

# Total probability of Red Sox winning the series
p_total <- p_win_3 + p_win_4 + p_win_5
p_total
```

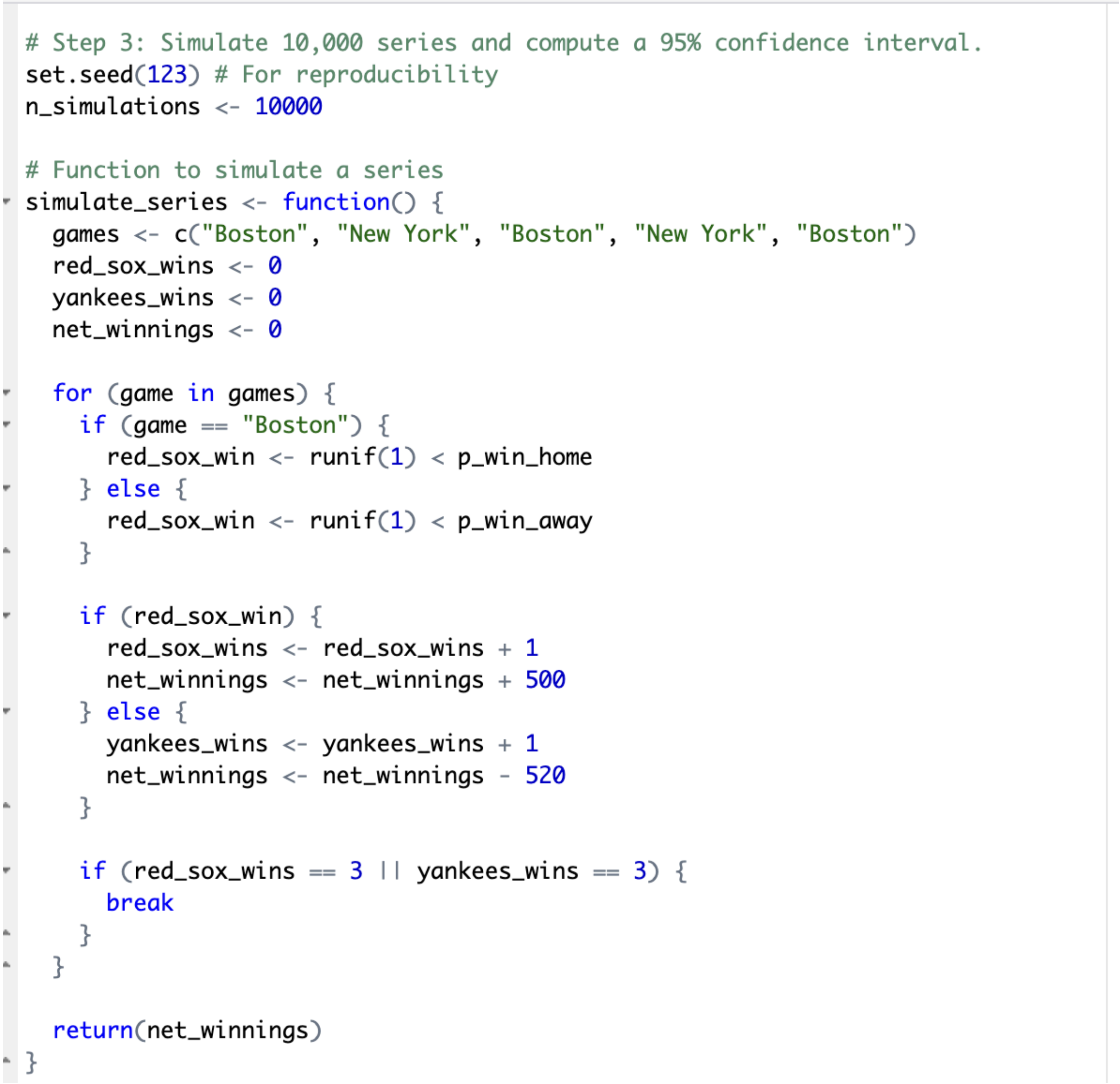


```
simula
simula
simula
```

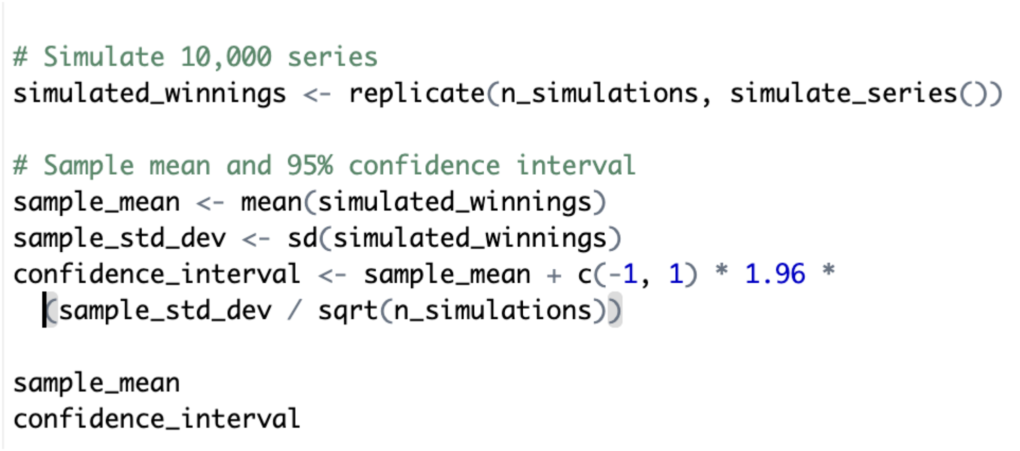
```
# Step 2: Construct the theoretical probability distribution for net winnings (X).
# Net winnings for each outcome
winnings_win_3 <- 1500
winnings_win_4 <- 980
winnings_win_5 <- 460
winnings_lose_3 <- -1560
winnings_lose_4 <- -1060
winnings_lose_5 <- -560

# Probabilities for each outcome
p_win_3
p_win_4
p_win_5

p_lose_3 <- (1 - p_win_home) * (1 - p_win_away) * (1 - p_win_home)
p_lose_3

p_lose_4 <- (p_win_home * (1 - p_win_away) * (1 - p_win_home) * (1 - p_win_away)) +
  ((1 - p_win_home) * p_win_away * (1 - p_win_home) * (1 - p_win_away))
p_lose_4

p_lose_5 <- (p_win_home * p_win_away * (1 - p_win_home) * (1 - p_win_away) * (1 - p_win
  (p_win_home * (1 - p_win_away) * p_win_home * (1 - p_win_away) * (1 - p_win_home)) +
  ((1 - p_win_home) * p_win_away * p_win_home * (1 - p_win_away) * (1 - p_win_home)) +
  ((1 - p_win_home) * p_win_away * (1 - p_win_home) * p_win_away * (1 - p_win_home))
p_lose_5

# Expected value (mean) of X
expected_value <- (winnings_win_3 * p_win_3) +
  (winnings_win_4 * p_win_4) +
  (winnings_win_5 * p_win_5) +
  (winnings_lose_3 * p_lose_3) +
  (winnings_lose_4 * p_lose_4) +
  (winnings_lose_5 * p_lose_5)
```



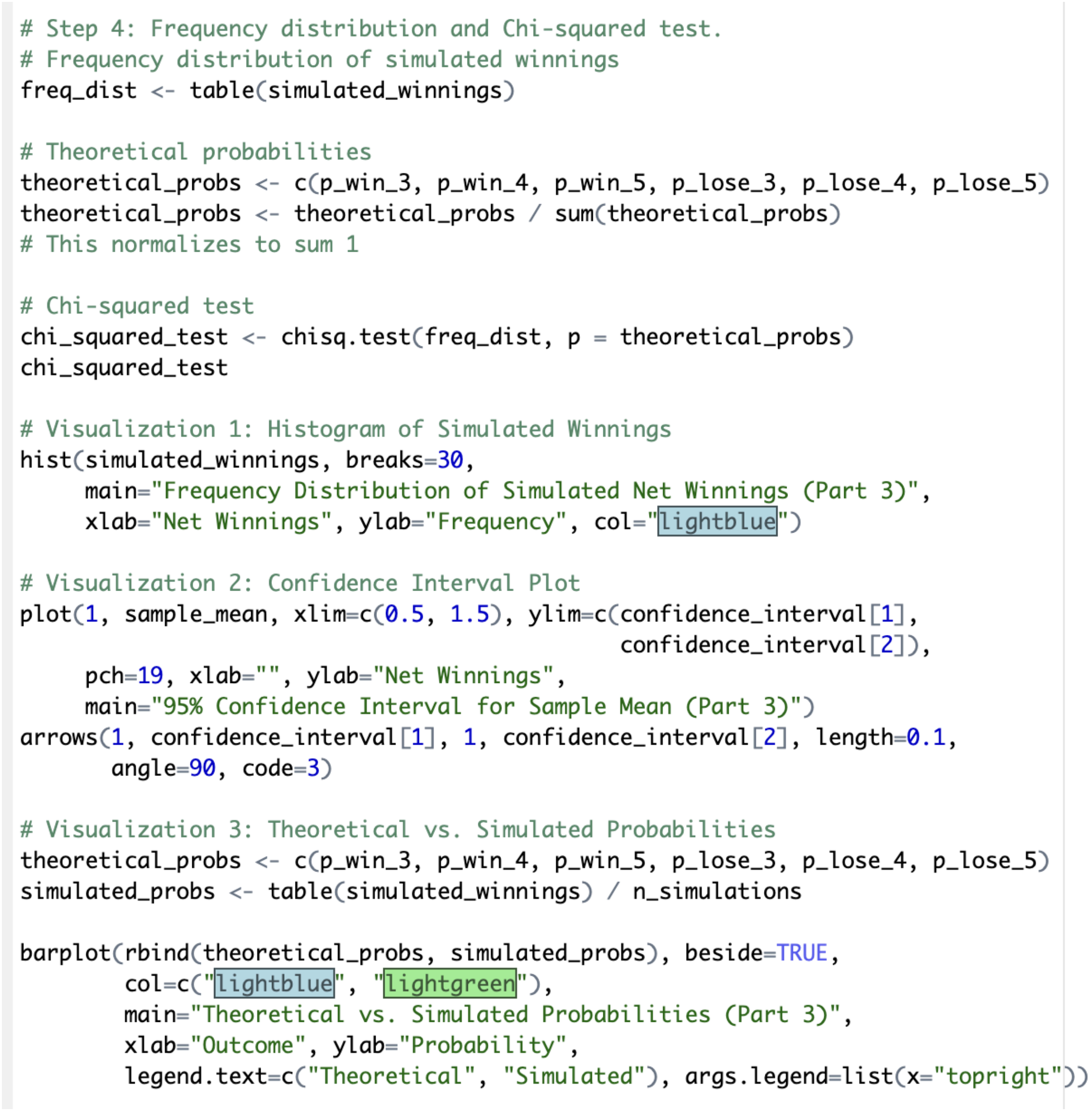
```
# Variance and standard deviation of X
variance <- ((winnings_win_3 - expected_value)^2 * p_win_3) +
  ((winnings_win_4 - expected_value)^2 * p_win_4) +
  ((winnings_win_5 - expected_value)^2 * p_win_5) +
  ((winnings_lose_3 - expected_value)^2 * p_lose_3) +
  ((winnings_lose_4 - expected_value)^2 * p_lose_4) +
  ((winnings_lose_5 - expected_value)^2 * p_lose_5)
std_dev <- sqrt(variance)

expected_value
std_dev
```



```r
# Step 3: Simulate 10,000 series and compute a 95% confidence interval.
set.seed(123) # For reproducibility
n_simulations <- 10000

# Function to simulate a series
simulate_series <- function() {
  games <- c("Boston", "New York", "Boston", "New York", "Boston")
  red_sox_wins <- 0
  yankees_wins <- 0
  net_winnings <- 0

  for (game in games) {
    if (game == "Boston") {
      red_sox_win <- runif(1) < p_win_home
    } else {
      red_sox_win <- runif(1) < p_win_away
    }

    if (red_sox_win) {
      red_sox_wins <- red_sox_wins + 1
      net_winnings <- net_winnings + 500
    } else {
      yankees_wins <- yankees_wins + 1
      net_winnings <- net_winnings - 520
    }

    if (red_sox_wins == 3 || yankees_wins == 3) {
      break
    }
  }

  return(net_winnings)
}
```



```r
# Simulate 10,000 series
simulated_winnings <- replicate(n_simulations, simulate_series())

# Sample mean and 95% confidence interval
sample_mean <- mean(simulated_winnings)
sample_std_dev <- sd(simulated_winnings)
confidence_interval <- sample_mean + c(-1, 1) * 1.96 *
  (sample_std_dev / sqrt(n_simulations))

sample_mean
confidence_interval
```



```
# Step 4: Frequency distribution and Chi-squared test.
# Frequency distribution of simulated winnings
freq_dist <- table(simulated_winnings)

# Theoretical probabilities
theoretical_probs <- c(p_win_3, p_win_4, p_win_5, p_lose_3, p_lose_4, p_lose_5)
theoretical_probs <- theoretical_probs / sum(theoretical_probs)
# This normalizes to sum 1

# Chi-squared test
chi_squared_test <- chisq.test(freq_dist, p = theoretical_probs)
chi_squared_test

# Visualization 1: Histogram of Simulated Winnings
hist(simulated_winnings, breaks=30,
     main="Frequency Distribution of Simulated Net Winnings (Part 3)",
     xlab="Net Winnings", ylab="Frequency", col="lightblue")

# Visualization 2: Confidence Interval Plot
plot(1, sample_mean, xlim=c(0.5, 1.5), ylim=c(confidence_interval[1],
                                             confidence_interval[2]),
     pch=19, xlab="", ylab="Net Winnings",
     main="95% Confidence Interval for Sample Mean (Part 3)")
arrows(1, confidence_interval[1], 1, confidence_interval[2], length=0.1,
       angle=90, code=3)

# Visualization 3: Theoretical vs. Simulated Probabilities
theoretical_probs <- c(p_win_3, p_win_4, p_win_5, p_lose_3, p_lose_4, p_lose_5)
simulated_probs <- table(simulated_winnings) / n_simulations

barplot(rbind(theoretical_probs, simulated_probs), beside=TRUE,
        col=c("lightblue", "lightgreen"),
        main="Theoretical vs. Simulated Probabilities (Part 3)",
        xlab="Outcome", ylab="Probability",
        legend.text=c("Theoretical", "Simulated"), args.legend=list(x="topright"))
```

Conclusion

This study evaluated the betting strategy for various series formats in baseball game. The best-of-three series with home advantage (Boston-New York-Boston) had an expected net gain of $57.89, but it was very risky. The New York-Boston-New-York format had negative expectations for -$31.24. The best-of-five format yielded the highest expected winnings of $152.89 with a high standard deviation of $954.27 and thus should be considered a risky strategy. The simulated outcomes confirmed theoretical expectations, validating the importance of home-field advantage and gaming sequences in determining betting profitability. Finally, although the best-of-five format appears most profitable, considerable uncertainty remains. A bettor's risk tolerance should indicate whether he should pursue this strategy, as the losses are steep despite a positive expected value.

References

Albright, S. (2016) Business Analytics. Sixth Edition. Cengage Learning. Boston, MA.

Evans, J. R. (2013). *Statistics, data analysis, and decision modeling: International Edition*. Pearson Higher Ed.