*Utilizes Excel to calculate probabilities, expected values, distributions, random variables, and Chi-Square tests.*

**Project**

**1**

P1

ALY6050 Intro to Enterprise Analytics

Project 1 – Analysis of Betting Strategy in Sports

**PREPERATION:**

By: John DiSessa

For: Professor Behboudi

On: February 27th, 2022

Introduction

This project analyzed various hypothetical 3 and 5 game series between the Red Sox and Yankees. It evaluates how effective the following betting strategy is. We win $500 for each Red Sox win and lose $520 for each Red Sox loss. Each game and each series is assumed to be independent, there can be no ties, a 3 game series ends when one team wins two games, and a 5 game series ends when one team wins three games. We assume that Red Sox have a 60% chance to win each game at home and a 43% chance to win on the road. For each part, we calculated the following:

* The Red Sox series win probability
* An expected value for betting for each series
* A 10,000 sample randomly generated simulation to test our betting strategy
* A 95% confidence interval to see the ranges of potential betting results
* A probability distribution, a frequency distribution, and a Chi Squared goodness of fit test to see how close each simulation matched the expected results

Part 1

For this part, there will be a 3-game series between the Red Sox and the Yankees, with game 1 being played at Fenway Park (home game for Red Sox), game 2 being played at Yankee Stadium (away game for the Red Sox), and game 3 back at Fenway park if it is needed. Before calculating the probability that the Red Sox will win the series, we need a list of all possible permutations for how the series will play out. There are 6 possible permutations listed in the first column. The probability of each permutation occurring is listed in the second column. The series result is listed in the third column. The Red Sox have a series win probability of .5664 (.258 + .2052 + .1032). The Red Sox are expected to win the series 56.64% of the time.

|  |  |  |
| --- | --- | --- |
| **Permutations** | **Series Probability** | **Series Result** |
| P(WW) = .6\*.43 | 0.258 | Win |
| P(WLW) = .6\*.57\*.6 | 0.2052 | Win |
| P(LWW) = .4\*.43\*.6 | 0.1032 | Win |
| P(LL) = .4\*.57 | 0.228 | Loss |
| P(LWL) = .4\*.43\*.4 | 0.0688 | Loss |
| P(WLL) = .6\*.57\*.4 | 0.1368 | Loss |

Before calculating our expected value for betting, we first need to find the probability of game 3 being played. Since the first team to win 2 games wins the series, there will always have to be at least 2 games played. However, the Red Sox or Yankees could win both games 1 and 2 and ‘sweep’ the other team so that game 3 is not played. The probability of game 3 being played is .514 (.2052 + .1032 + .0688 + .1368). To calculate our expected value based on our betting strategy, we added up the expected values of each game. Games 1 and 2 expected values are calculated by adding 500\*win probability and -520\*lose probability. However for game 3, we need to do the same except multiply it by the probability of game 3 even occurring. With our current betting strategy, we have an expected value of $57.89.

|  |  |  |
| --- | --- | --- |
| **Game 1** | **Game 2** | **Game 3\* if necessary** |
| Fenway | Yankee Stadium | Fenway |
| (.6\*500)+(.4\*-520) | (.43\*500)+(.57\*-520) | ((.6\*500)+(.4\*-520))\*.514 |
| $92 | -$81.40 | $47.29 |

Even though our expected value is positive, there is still a large amount of variability since the Red Sox only have a 56.64% chance to win. If the Red Sox win the series in 2 games, we would win $1,000. If the Red Sox win the series in 3 games, we would win $480. If the Red Sox lose the series in 2 games, we would lose $1,040. If the Red Sox lose the series in 3 games, we would lose $540. To measure the variation, we calculated the standard deviation of our expected value. The variance of games 1-3 are 249696, 255002, and 255924 respectively. Our standard deviation is $872.14.

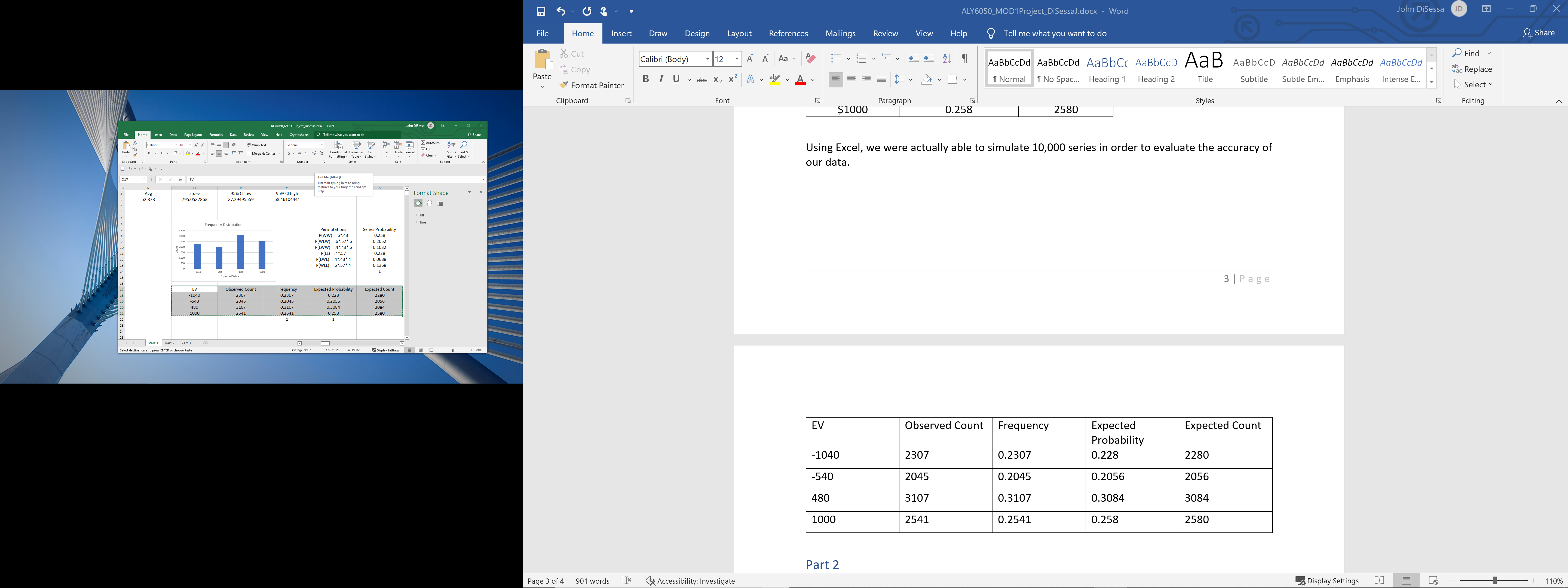
Now that we know there are 4 possible betting outcomes, we can find the probability of each outcome occurring. The $1000 probability is the same as the Red Sox winning in 2 games. The -$1040 probability is the same as the Red Sox losing the series in 2 games. The $480 probability is the sum of all permutations of the Red Sox winning in 3 games. The -$540 probability is the sum of all permutation of the Red Sox losing in 3 games. All of these probabilities add up to 1 so we know we have all possible permutations accounted for. The expected count is the number of times we would have each expected value of we used our betting strategy for 10,000 independent series between the Red Sox and Yankees.

|  |  |  |
| --- | --- | --- |
| **Expected Value** | **Expected Probability** | **Expected Count** |
| -$1040 | 0.228 | 2280 |
| -$540 | 0.2056 | 2056 |
| $480 | 0.3084 | 3084 |
| $1000 | 0.258 | 2580 |

Using Excel, we were actually able to simulate 10,000 series in order to evaluate the accuracy of our data. To set up the simulation for each series, game 1 had a 60% chance of $500 and a 40% chance of -$520. Game 2 had a 43% chance of $500 and a 57% chance of -$520. Game 3 had the same odds for game 1, except that $0 was returned if games 1 and 2 were either win/win or lose/lose. We then averaged all 10,000 series betting results and got an expected value of $52.88 with a standard deviation of 795. Even though that is a little lower that our expected value of $57.89, we are still within our 95% confidence interval of (37.29, 68.46).

Now that we have our simulation results, we can create a frequency distribution table and chart and use a Chi-Square Goodness of Fit Test to see how close our distribution was to our expected distribution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **EV** | **Observed Count** | **Frequency** | **Expected Probability** | **Expected Count** |
| -1040 | 2307 | 0.2307 | 0.228 | 2280 |
| -540 | 2045 | 0.2045 | 0.2056 | 2056 |
| 480 | 3107 | 0.3107 | 0.3084 | 3084 |
| 1000 | 2541 | 0.2541 | 0.258 | 2580 |



To conduct our Chi Square test with α = .05, we used the following null and alternative hypotheses:

* H0: No significant difference between our observed and expected values
* Ha: There is a significant difference between our observed and expected values

Our Chi-Square value was 1.14 and our p-value was .77. Since our p-value is greater than our cutoff point (.05). We fail to reject the null hypothesis. If we cannot reject our null, we accept that there is not a significant difference between our observed and expected values. Therefore, our simulation distribution is similar to our expected distribution with 95% certainty.

Part 2

For Part 2, we conducted the exact same analysis and kept the same betting strategy, except the order of the series has changed. Now, the Yankees play at home for games 1 and 3 (if necessary) and the Red Sox only play at home for game 2. The Red Sox still have a 60% chance to win at home and a 43% chance to win at Yankee Stadium, however the Red Sox will only get one home game this series instead of 1 or 2 per series. There are still the same 6 permutation of how the series will play out, however the probability of some permutations occurring have changed because the Red Sox have less home games than in Part 1. Winning or losing the series in 2 games hasn’t changed because each team gets 1 home game, just like in part one even though the order of home/away has changed.

|  |  |  |
| --- | --- | --- |
| **Permutations** | **Series Probability** | **Series Result** |
| P(WW) = .43\*.6 | 0.258 | Win |
| P(WLW) = .43\*.4\*.43 | 0.07396 | Win |
| P(LWW) = .57\*.6\*.43 | 0.14706 | Win |
| P(LL) = .57\*.4 | 0.228 | Loss |
| P(LWL) = .57\*.6\*.57 | 0.19494 | Loss |
| P(WLL) = .43\*.4\*.57 | 0.09804 | Loss |

Their probability of winning each series is now .479 or 47.9% and there is still a probability of .514 that a game 3 will be played.

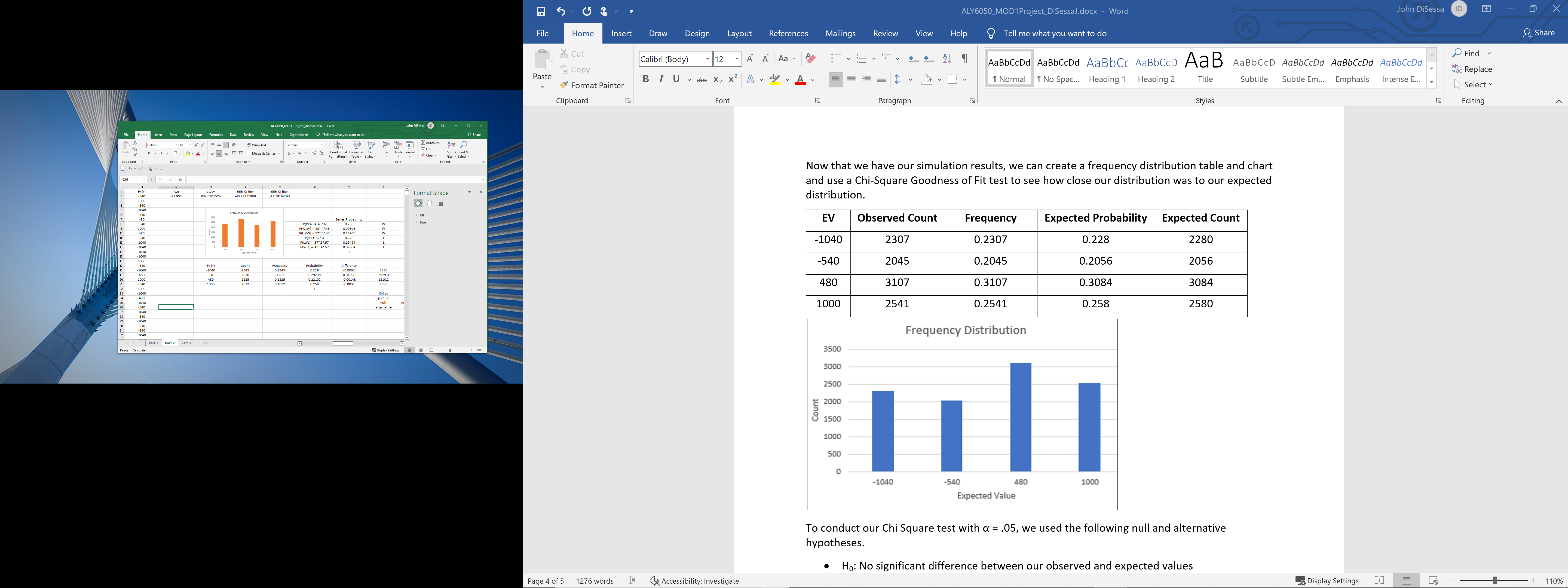
|  |  |  |
| --- | --- | --- |
| **Game 1** | **Game 2** | **Game 3\* if necessary** |
| Yankee Stadium | Fenway | Yankee Stadium |
| (.43\*500)+(.57\*-520) | (.6\*500)+(.4\*-520) | ((.43\*500)+(.57\*-520))\*.514 |
| -$81.40 | $92 | -$41.84 |

Our expected value for these series is -$31.24 with a standard deviation of 874. Our 95% confidence interval is (-43.72, -12.18). Since the home team order has changed, it is now more likely than before that we will lose $540 than win $480. There used to be a 20.56% chance of losing $540, now it is a 29.30%. There used to be a 30.84% chance of winning $480, now it is down to 22.1%.

|  |  |  |
| --- | --- | --- |
| **Expected Value** | **Expected Probability** | **Expected Count** |
| -$1040 | 0.228 | 2280 |
| -$540 | 0.29298 | 2930 |
| $480 | 0.22102 | 2210 |
| $1000 | 0.258 | 2580 |

We conducted another 10,000 simulations like before except with the updated probabilities for the new series.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Expected Value** | **Observed Count** | **Frequency** | **Expected Probability** | **Expected Count** |
| -1040 | 2343 | 0.2343 | 0.228 | 2280 |
| -540 | 2820 | 0.282 | 0.29298 | 2929.8 |
| 480 | 2225 | 0.2225 | 0.22102 | 2210.2 |
| 1000 | 2612 | 0.2612 | 0.258 | 2580 |



To conduct our Chi Square test with α = .05, we used the following null and alternative hypotheses.

* H0: No significant difference between our observed and expected values
* Ha: There is a significant difference between our observed and expected values

Our Chi-Square value was 6.35 and our p-value was .10. Since our p-value is greater than our cutoff point (.05). We fail to reject the null hypothesis. If we cannot reject our null, we accept that there is not a significant difference between our observed and expected values. Therefore, like in Part 1, our simulation distribution is similar to our expected distribution with 95% certainty.

Part 3

In this scenario, we now have 5-game series instead of 3-game series. The Red Sox will be the home team for games 1,3 and 5. The Yankees will be the home team for games 2 and 4. The Red Sox still have a 60% and 43% chance to win each game, however we now have 20 possible permutations for this series instead of 6. The series will end when one team wins 3 games so the series could end in 3,4, or 5 games.

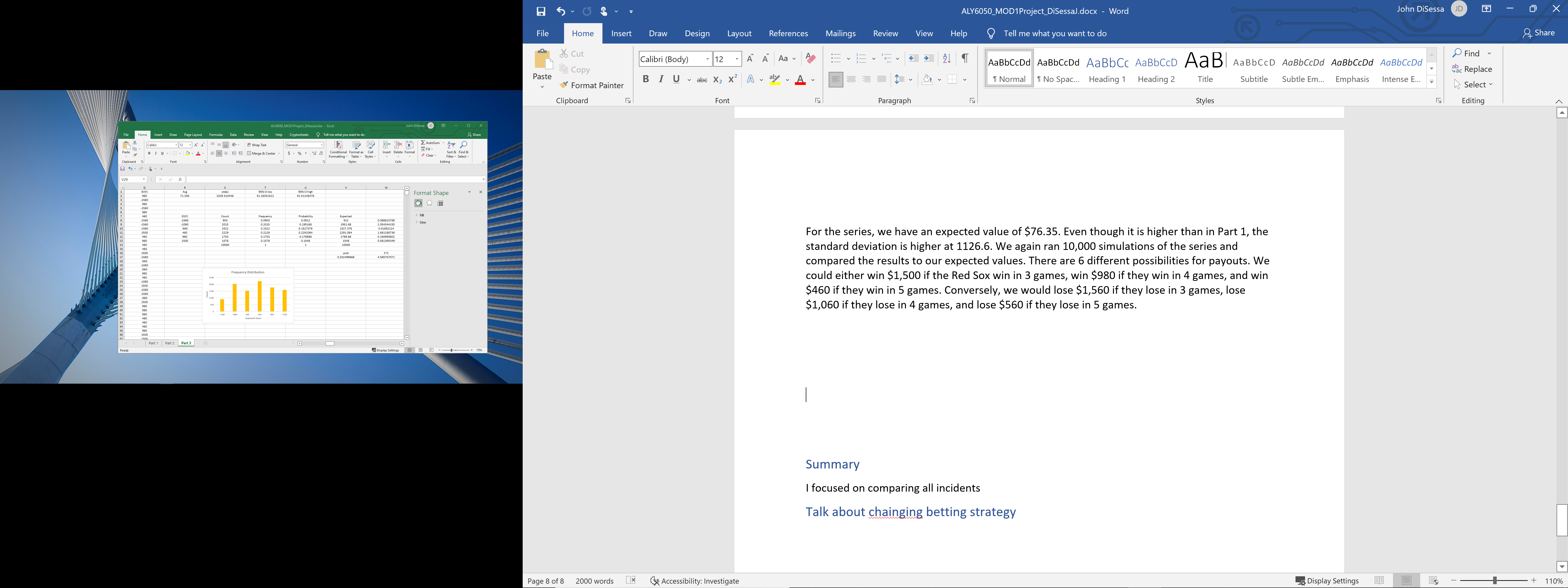
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Permutations** | **Math** | **Series Probability** | **Net Result** | **Series Result** |
| P(WWW) | 0.6\*0.43\*0.6 | 0.1548 | $1500 | Win |
| P(WWLW) | 0.6\*0.43\*0.4\*0.43 | 0.044376 | $980 | Win |
| P(WLWW) | 0.6\*0.57\*0.6\*0.43 | 0.088236 | $980 | Win |
| P(LWWW) | 0.4\*0.43\*0.6\*0.43 | 0.044376 | $980 | Win |
| P(WWLLW) | 0.6\*0.43\*0.4\*0.57\*0.6 | 0.0352944 | $460 | Win |
| P(WLLWW) | 0.6\*0.57\*0.4\*0.43\*0.6 | 0.0352944 | $460 | Win |
| P(WLWLW) | 0.6\*0.57\*0.6\*0.57\*0.6 | 0.0701784 | $460 | Win |
| P(LLWWW) | 0.4\*0.57\*0.6\*0.43\*0.6 | 0.0352944 | $460 | Win |
| P(LWWLW) | 0.4\*0.43\*0.6\*0.57\*0.6 | 0.0352944 | $460 | Win |
| P(LWLWW) | 0.4\*0.43\*0.4\*0.43\*0.6 | 0.0177504 | $460 | Win |
| P(LLL) | 0.4\*0.57\*0.4 | 0.0912 | -$1560 | Loss |
| P(LLWL) | 0.4\*0.57\*0.6\*0.57 | 0.077976 | -$1060 | Loss |
| P(LWLL) | 0.4\*0.43\*0.4\*0.57 | 0.039216 | -$1060 | Loss |
| P(WLLL) | 0.6\*0.57\*0.4\*0.57 | 0.077976 | -$1060 | Loss |
| P(LLWWL) | 0.4\*0.57\*0.6\*0.43\*0.4 | 0.0235296 | -$560 | Loss |
| P(LWWLL) | 0.4\*0.43\*0.6\*0.57\*0.4 | 0.0235296 | -$560 | Loss |
| P(LWLWL) | 0.4\*0.43\*0.4\*0.43\*0.4 | 0.0118336 | -$560 | Loss |
| P(WWLLL) | 0.6\*0.43\*0.4\*0.57\*0.4 | 0.0235296 | -$560 | Loss |
| P(WLLWL) | 0.6\*0.57\*0.4\*0.43\*0.4 | 0.0235296 | -$560 | Loss |
| P(WLWLL) | 0.6\*0.57\*0.6\*0.57\*0.4 | 0.0467856 | -$560 | Loss |

According to the 10 permutations where the Red Sox win, they have a 56.09% chance to win the series. Before calculating our expected value for betting these series with the same betting strategy as before, we need to determine the probability of games 4 and 5 being played. Based on the chart above, game 4 has a 75.4% chance of happening and game 5 has a 38.13% chance of happening.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Game 1** | **Game 2** | **Game 3** | **Game 4 \*\*** | **Game 5 \*\*** |
| Fenway | Yankee Stadium | Fenway | Yankee Stadium | Fenway |
| (.6\*500)+(.4\*-520) | (.43\*500)+(.57\*-520) | (.6\*500)+(.4\*-520) | ((.43\*500)+(.57\*-520))\*.754 | ((.6\*500)+(.4\*-520))\*.3813 |
| $92 | -$81.4 | $92 | -$61.38 | $35.13 |

For the series, we have an expected value of $76.35. Even though it is higher than in Part 1, the standard deviation is higher at 1127. We again ran 10,000 simulations of the series and compared the results to our expected values. There are 6 different possibilities for payouts. We could either win $1,500 if the Red Sox win in 3 games, win $980 if they win in 4 games, or win $460 if they win in 5 games. Conversely, we would lose $1,560 if they lose in 3 games, lose $1,060 if they lose in 4 games, or lose $560 if they lose in 5 games. Our simulation had an expected value of $71.34, which is well within our 95% confidence interval (51.16, 91.51), and a standard deviation of 1029.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **EV** | **Observed Count** | **Frequency** | **Expected Probability** | **Expected Count** |
| -1560 | 903 | 0.0903 | 0.0912 | 912 |
| -1060 | 2015 | 0.2015 | 0.195168 | 1951.68 |
| -560 | 1522 | 0.1522 | 0.1527376 | 1527.376 |
| 460 | 2229 | 0.2229 | 0.2291064 | 2291.064 |
| 980 | 1753 | 0.1753 | 0.176988 | 1769.88 |
| 1500 | 1578 | 0.1578 | 0.1548 | 1548 |



Like in Parts 1 and 2, we had to conduct a Chi-Square Goodness of Fit Test to see how close our distribution was to our expected distribution. To conduct our Chi Square test with α = .05, we used the following null and alternative hypotheses.

* H0: No significant difference between our observed and expected values
* Ha: There is a significant difference between our observed and expected values

Our Chi-Square value was 4.59 and our p-value was .33. Since our p-value is greater than our cutoff point (.05). We fail to reject the null hypothesis. If we cannot reject our null, we accept that there is not a significant difference between our observed and expected values. Therefore, like in Parts 1 and 2, our simulation distribution is similar to our expected distribution with 95% certainty.

Summary

The main takeaway from our analysis is that home field advantage significantly affects the outcomes of the series and our betting strategy. Both the Red Sox and Yankees have relatively large home field advantages so it is not surprising to see the series odds and expected values change so much between Parts 1-3. Our expected values are correct and were verified by our 10,000 simulations for each part, so we feel confident commenting on the betting strategy. One simple improvement would be to only bet on series where the Red Sox have the home field advantage, since those series had a positive expected value and Part 2 did not. However, we can improve it even further in a simple way, by only betting on games where the Red Sox are the home team. For Red Sox home games, we have an expected value of $92, which is larger than the series expected value by betting every game. We could also use this model to try other scenarios, such as if the Red Sox had a 55% chance to win at home. By reducing their home field advantage, we could change our expected values game by game as the series progresses. In conclusion, we demonstrated our abilities to calculate probabilities, expected values, distributions, random variables, and Chi-Square tests. We can properly simulate scenarios and evaluate those simulated experiments on their fit to the expected results. We can use these formulas and strategies and apply them to other sports betting scenarios.