Probability and Randomness of the Powerball Lottery

Aldo Madrid

CIS 320

Introduction:

The winning numbers of the lottery are expected to be completely random. If this is the case there is no way to predict them. However, if there is a correlation between the numbers that are drawn and the date that those numbers are drawn then there would be a way to predict them. Also, if there is a way to predict, to a certain extent, the numbers that will be drawn in the future from the information of the numbers that have been drawn in the past then there is a chance to have an advantage over other players of the lottery. This experiment will be focusing on the numbers that tend to be drawn more often than other numbers in Powerball lottery draws. The experiment will be analyzing the data of drawings that happened between the years 2010 to 2015. In October 7, 2015 the rules of the Powerball changed and more numbers were added to each drawing. The experiment will also only focus on the drawing of the "Five Winners," that is the numbers other than the power ball. The reason for this is that the numbers in the power balls is different than the numbers in the "Five Winners." Therefore, an attempt to predict the "next" lottery number drawing will not be made. However, an attempt at replicating the data gathered from the "Five Winners" between the years 2010 and 2015 will be made with the program R. Thus, the results of this experiment can be used as a stepping stone towards a precise and accurate simulation and prediction of the "Five Winners."

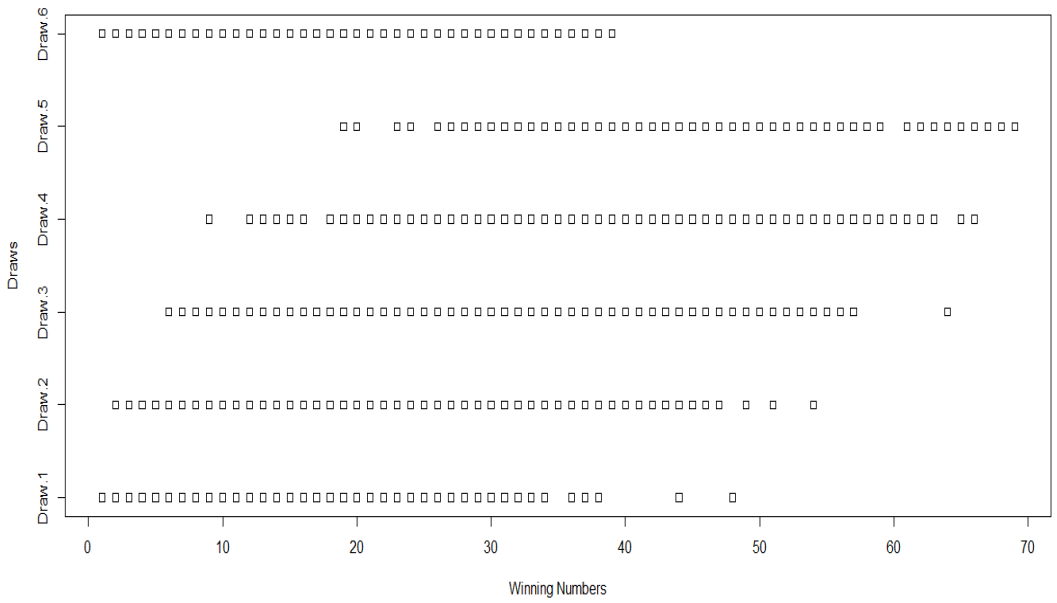
Procedure:

Obtaining the winning numbers of the Powerball:

1. The data set that was used for this experiment can be acquired at Data.gov.
2. This data set contains the winning numbers from 2010 to 2016 draws, however, only the draws up to the date of October 7, 2015 were used.

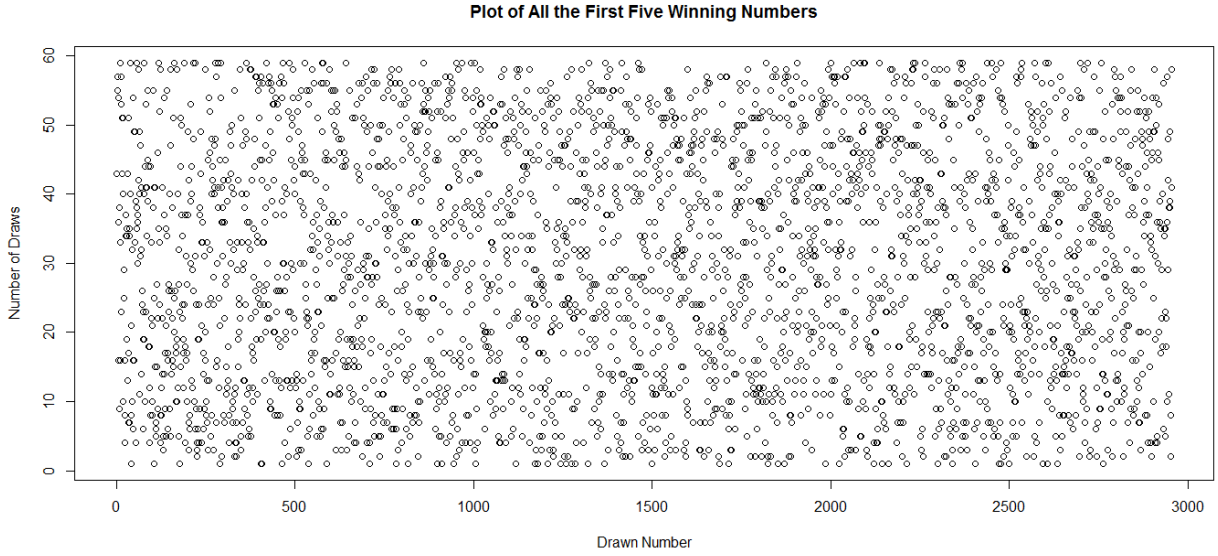
Analyzing the winning numbers of the Powerball:

1. The first step towards analyzing the data was to place it into data frame for easy access and for easy analysis.
2. Producing a strip chart in R provided useful information about the data.



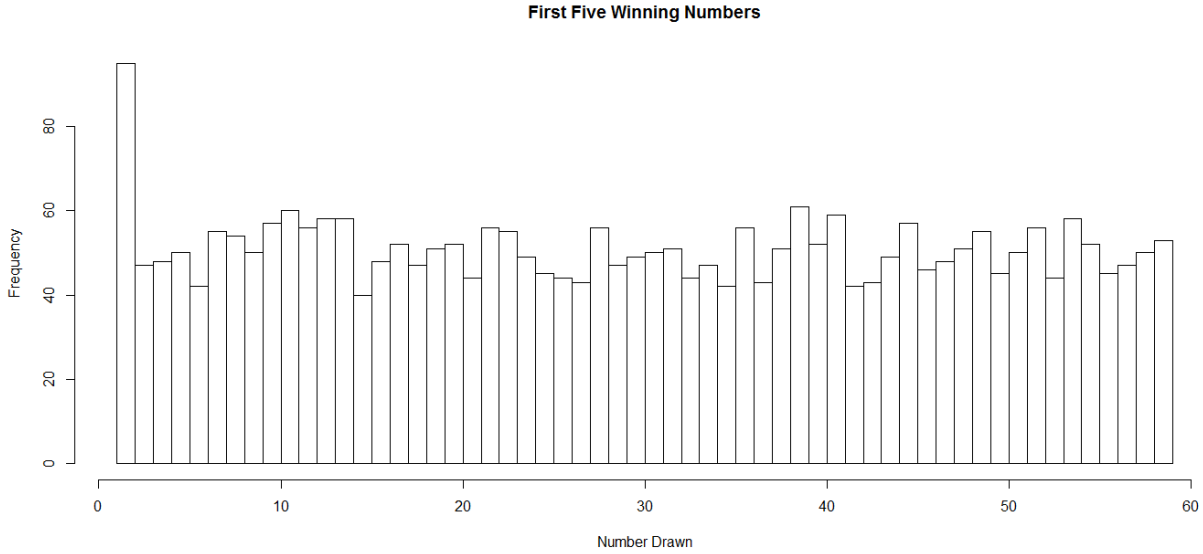
It can be seen that the data for each drawing was previously ordered, with “Draw 5” having the highest numbers for each draw. “Draw 6” represents the Powerball since it seems to have a uniform distribution between the numbers 1 through 39. Although the data seems to be pre-ordered, this chart shows that the first five draws tend to include as many low numbers as they have high numbers. In other words, the data of these draws seems to be uniform.

3. Producing a plot of this list of number further shows the insights about the distribution of these numbers. Since the numbers were placed in order they had to be randomized with R before plotting.



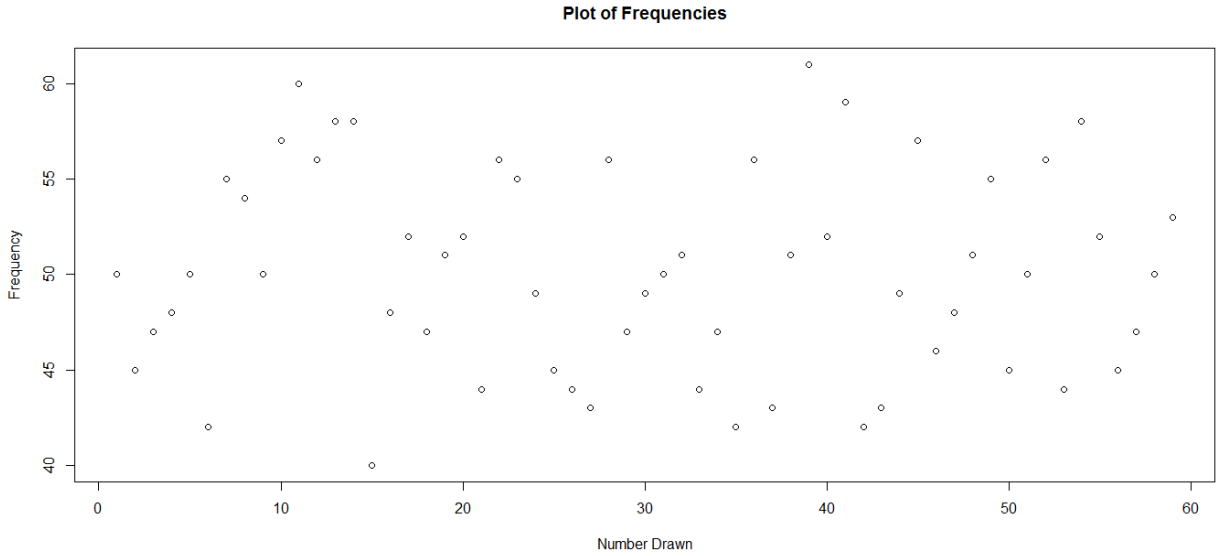
This plot shows a uniform distribution of the first five draw numbers.

4. To learn about the distribution of these numbers a histogram was produced.



The winning five numbers seem to have a fairly uniform distribution over all. This is a sign of true randomness. However, there are a few numbers that seem to have a significant difference in frequencies, such as the number 1 and the number 38.

5. To make sure that the histogram is showing correct information a plot of frequencies was made.



As it can be seen in this plot, the histogram showed some invalid information. We can now see that is not an outlier. Also, we can see that the number 39 is the one that tends to come up the most. To validate our findings, a table of the data was produced.

```
tDraws
31 51 58 24 30 44  4 16 47  3 18 29 34 57 46  2 25 50 56 21 26 33 53 27 37 43  6 35 42 15
50 50 50 49 49 49 48 48 48 47 47 47 47 47 46 45 45 45 45 44 44 44 44 43 43 43 42 42 42 40
> |
```

```
tDraws
39 11 41 13 14 54 10 45 12 22 28 36 52  7 23 49  8 59 17 20 40 55 19 32 38 48  1  5  9 31
61 60 59 58 58 58 57 57 56 56 56 56 56 55 55 55 54 53 52 52 52 52 51 51 51 51 50 50 50 50
```

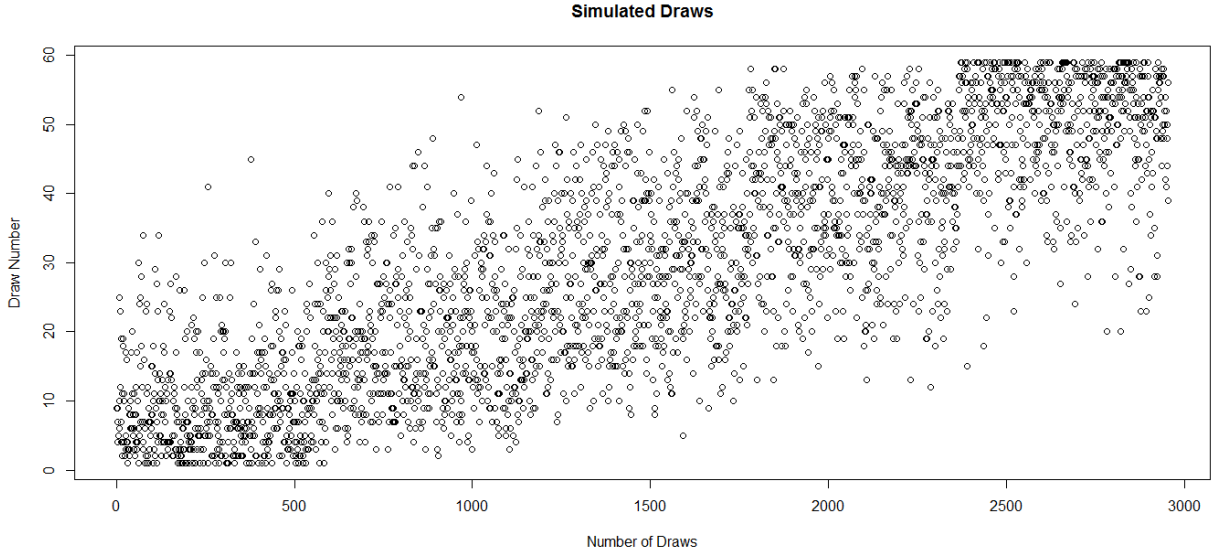
Creating a table of the frequencies can tell us that indeed the number 39 is the one that shows up the most and that 1 is somewhere in the middle in terms of frequencies.

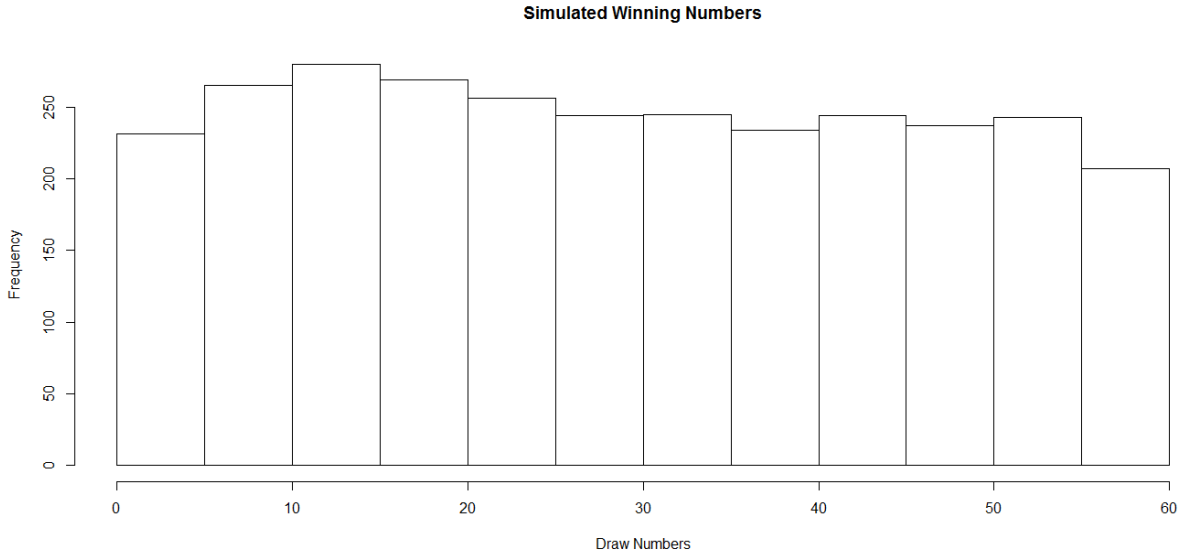
Reproducing the data:

1. To randomly generate data to resemble that of the lottery winning numbers the probabilities of each number were first calculated using the following formula:

   Probability = (Frequency) / (Total number of draws).

   This formula was then applied to each winning number.
2. With these probabilities, 591 draws of 5 numbers were simulated using the sample() function in R. The results were as follows:





The results are similar to those of the actual drawings. However, the histogram shows that the simulated data tends to have a bit more of a normal distribution than the actual data.

Discussion:

Explaining the results:

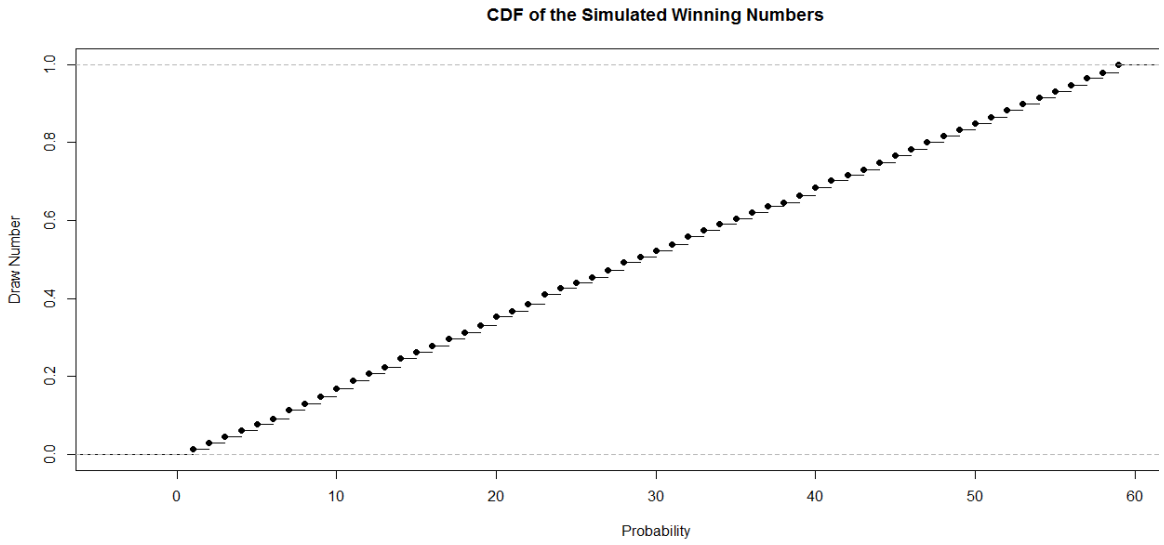
Even though the results of the draws of the lottery numbers are expected to be completely random, they are not. The results had some tendencies. By taking a look at a table of frequencies, one can spot those numbers that tend to be drawn more often than others:

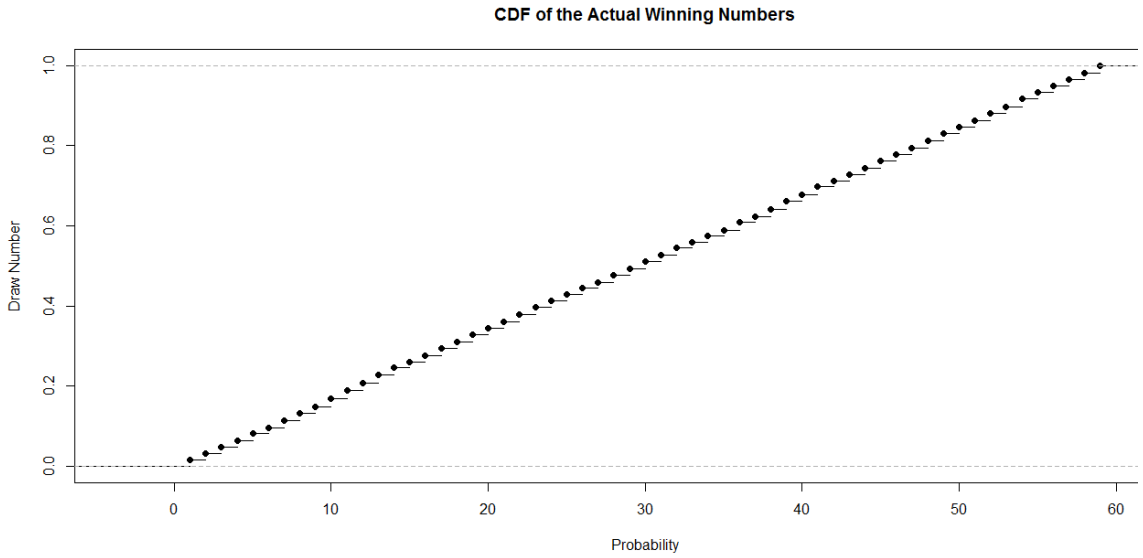
```
> head(sort(table(tDraws)), 30)
tDraws
15  6 35 42 27 37 43 21 26 33 53  2 25 50 56 46  3 18 29 34 57  4 16 47 24 30 44  1  5  9
40 42 42 42 43 43 43 44 44 44 44 45 45 45 45 46 47 47 47 47 47 48 48 48 49 49 49 50 50 50
> tail(sort(table(tDraws)), 30)
tDraws
 9 31 51 58 19 32 38 48 17 20 40 55 59  8  7 23 49 12 22 28 36 52 10 45 13 14 54 41 11 39
50 50 50 50 51 51 51 51 52 52 52 52 53 54 55 55 55 56 56 56 56 56 57 57 58 58 58 59 60 61
```

As it can be seen, 39 has been drawn 61 times while 15 only 40 times.

Simulating the draws can provide us with some insights that the actual data cannot. We can simulate many draws to learn about the numbers that tend to appear more often. However, to take advantage of these insights though, the data produced by the simulation has to first be considered an accurate representation of the actual data. Taking a look at a couple of analyzes of both data can help us make this decision.

1. Plotting the cumulative distribution function for both sets of data (function ecdf() in R was used).





From these CDF's it can be seen that both sets of data have constant probabilities, thus they are uniform random variables.

2. Finally, the fairness of the simulation was analyzed. It is known that the lottery is considered a fair game, otherwise the government wouldn't allow it continue. The fairness of both games was tested with a Chi-squared test for fairness. Here are the results.

<u>Chi-squared Test Results for Actual Lottery Winning Numbers</u>

```
> chisq.test(lottoNumbersFreq, p = prob)

        Chi-squared test for given probabilities

data:  lottoNumbersFreq
X-squared = 32.956, df = 58, p-value = 0.9967
```

As expected, the p-value is very high. The drawing of the “five winning” numbers is fair.

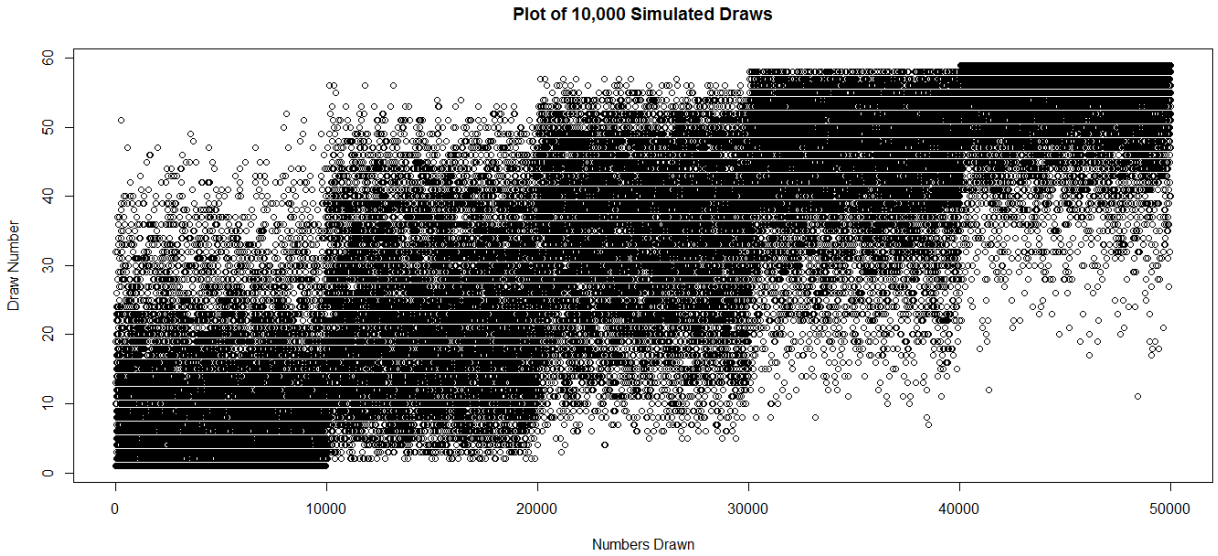
Chi-squared Test Results for Actual Lottery Winning Numbers

```
> chisq.test(simDrawNumFreq, p=prob)

        Chi-squared test for given probabilities

data:  simDrawNumFreq
X-squared = 76.562, df = 58, p-value = 0.05174
```

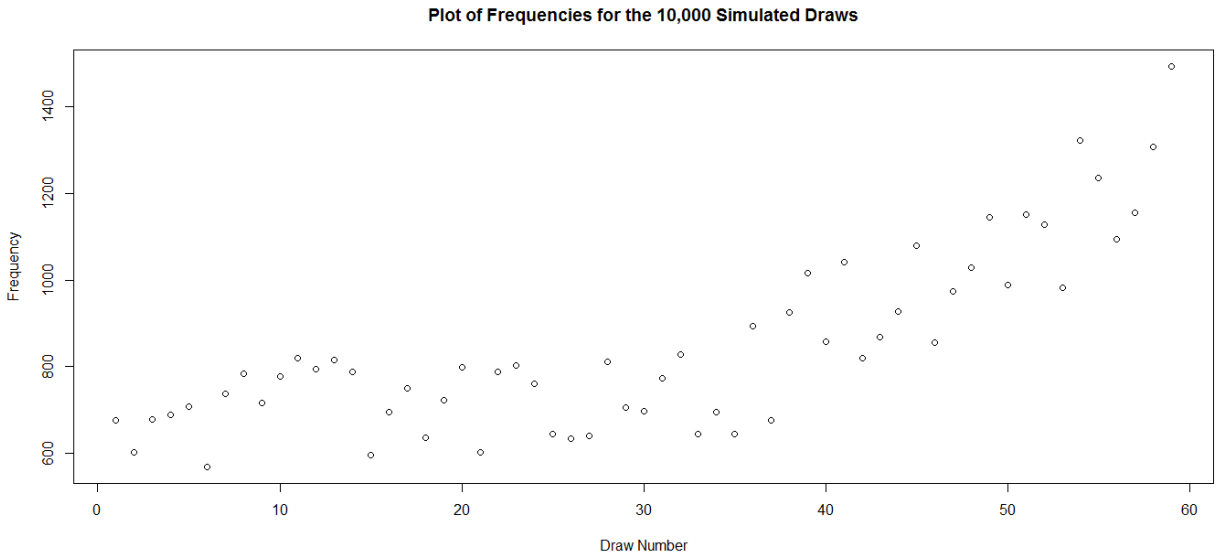
The p-value for this test is a lot lower than the previous one. However, this p-value is still above the threshold for rejecting the null hypothesis. Thus, the simulated winning numbers can be considered “fair.”

After knowing that our simulation can be considered a “fair game” and that it has the same kind of CDF as the actual lottery number drawings, a large number of drawings can be simulated to exploit the tendencies in our favor.

After simulating 10, 000 draws we can see that there are a few ranges of numbers that tend to come up more often.



The tendencies of the winning numbers are more prominent in this plot. However, this plot only tells us once again what we already knew, a draw tends to have as many low numbers as it has big numbers.



A plot of the frequencies can tell us a bit more about the numbers that tend to come up the most. In this plot it can be seen that the high numbers tend to come up, significantly, more than the lower numbers. This is a finding to consider.

Conclusion:

Predicting the next outcome from a roll of a die is easy. To be right about the outcome is difficult, however, especially if the die is a fair die. An advantage over predicting the outcome of the die roll can be obtained only if the die is an unfair die and the numbers that tend to come up the most are known. The same applies to the lottery winning numbers. The outcomes might be many but if we know that certain numbers tend to appear more often than others than our chances of predicting the next outcome improves. In this experiment it was noted that higher numbers tend to appear more often than lower numbers in our simulation. This information can certainly be useful. However, the rules of the Powerball have recently changed. This information can only support the idea that the numbers draws in the lottery have some tendencies and that these tendencies can be found using data analysis.

Code:

```
##########Code
#Reading and transforming the data
lottoNums <- read.csv('https://data.ny.gov/api/views/d6yy-54nr/rows.csv?accessType=DOWNLOAD')
lottoNums <- lottoNums$Winning.Numbers
lottoNums <- as.vector(lottoNums)
lottoNums <- strsplit(lottoNums, " ")
lottoNums <- unlist(lottoNums)
lottoNums <- as.numeric(lottoNums)
drawNames <- c("Draw 1", "Draw 2", "Draw 3", "Draw 4", "Draw 5", "Draw 6")
list.lottoNums <- split(lottoNums, drawNames)
df.lottoNums <- as.data.frame(list.lottoNums)
#analyzing the data
#separating draws
first5Draws <- c("Draw.1", "Draw.2", "Draw.3", "Draw.4", "Draw.5")
fDraws <- df.lottoNums[first5Draws]
draw1 <- fDraws[["Draw.1"]]
draw2 <- fDraws[["Draw.2"]]
draw3 <- fDraws[["Draw.3"]]
draw4 <- fDraws[["Draw.4"]]
draw5 <- fDraws[["Draw.5"]]
#creating training data
tDraw1 <- draw1[43:633] #powerlotto rules changed on October 7, 2015
tDraw2 <- draw2[43:633]
tDraw3 <- draw3[43:633]
tDraw4 <- draw4[43:633]
tDraw5 <- draw5[43:633]
#putting all the data together in a list
tDraws <- append(tDraw1, tDraw2)
```

```
tDraws <- append(tDraws, tDraw3)
tDraws <- append(tDraws, tDraw4)
tDraws <- append(tDraws, tDraw5)
#creating a plot and a histogram with the data
plot(tDraws)
hist(tDraws, breaks = 59)
#creating sample data with corresponding probabilities
#obtaining frequencies
library(plyr)                   #loading library "plyr" to make use of count()
tDrawTable <- count(tDraws)
tDrawFreq <- tDrawTable$freq
tDrawNums <- tDrawTable$x
#obtaining probabilities
tDrawProb = tDrawFreq/sum(tDrawFreq)          #sum(tDrawFreq) is the number of draws
sum(tDrawProb)                          #making sure all probabilities add up to one
#ploting probabilities
xRange <- seq(1, 59, by = 1)
plot(xRange, tDrawProb, main = "Plot of Probabilities", xlab = "Draw Numbers", ylab = "Probabilities")
#create sample data
set.seed(100)
sDraw <- c()
for(i in 1:length(tDraw1)){                   #length(tDraw1) is the number of draws
        x <- sample(tDrawNums, 5, replace = FALSE, prob = tDrawProb)
        x <- sort(x)
        x <- list(x)
        sDraw[i] <- x
}
#separating sample data
sDraw <- unlist(sDraw)
```

```
sDraw <- split(sDraw, first5Draws)
sDraw1 <- sDraw[["Draw.1"]]
sDraw2 <- sDraw[["Draw.2"]]
sDraw3 <- sDraw[["Draw.3"]]
sDraw4 <- sDraw[["Draw.4"]]
sDraw5 <- sDraw[["Draw.5"]]
#combining all numers in a string
sDraws <- append(sDraw1, sDraw2)
sDraws <- append(sDraws, sDraw3)
sDraws <- append(sDraws, sDraw4)
sDraws <- append(sDraws, sDraw5)
#analyzing sample data
plot(sDraws)
hist(sDraws, breaks = 59)
```