Assignment 1 – Golf Ball Statistics

Yibei Jiang

02/07/2020

I. Introduction

In this homework, we have the following scenario:

> Allan Rossman used to live along a golf course and collected the golf balls that landed in his yard. Most of these golf balls had a number on them. Allan tallied the numbers on the first 500 golf balls that landed in his yard one summer.
>
> Specifically, he collected the following data:
>
> - 138 golf balls numbered 2
> - 107 golf balls numbered 3
> - 104 golf balls numbered 4
> - 14 "others"

We are interested in the following:

Question: What is the distribution of these numbers?

II. Methods

0. Test statistics

In order to investigate the distribution of the golf ball numbers, we can test if the golf ball numbers, 1, 2, 3, 4 have equal chance of occurring. In this report, a simulation-based hypothesis test is conducted. The test statistics chosen in this report are:

- The minimum frequency of golf ball numbers
- The maximum frequency of golf ball numbers
- The variance of frequency among golf ball numbers

The null hypothesis: all numbers 1, 2, 3, 4 are equally distributed.

The alternative hypothesis: the numbers 1, 2, 3, 4 are not equally distributed.

Though we can use chi-square test to test whether the numbers are evenly distributed, we will first focus on using simulation to test for the significance of the associated test statistics and compare the results with that of chi-square test.

1. Simulation

First, 10000 simulations for each test statistic are created to find out their behavior under null hypothesis. In each simulation, we used a random number generator to sample 486 numbers among integers 1 through 4. These numbers follow a discrete uniform distribution according to the null hypothesis. Then, all 3 test statistics are calculated for the simulated data.

```r
# null hypothesis: all the numbers are equally likely
set.seed(123)
NumberOfSims <- 10000
DataSize <- 486

golfMinFreq <- rep(0,NumberOfSims) # The minimum frequency of golf ball numbers
golfMaxFreq <- rep(0,NumberOfSims) # The maximum frequency of golf ball numbers
golfVar<- rep(0,NumberOfSims) # The variance of frequency among golf ball numbers

# Generate test statistics under null hypothesis
for (sim in 1:NumberOfSims){
  randomNumbers <- sample(1:4, DataSize,replace = TRUE)
  golfMinFreq[sim] <- min(table(randomNumbers))
  golfMaxFreq[sim] <- max(table(randomNumbers))
  golfVar[sim] <- var(table(randomNumbers))
}
```

Here is a table showing the first few entries in the simulated data. The table has 10000 rows and 3 columns. The rows correspond to the number of simulation and each column corresponds to the minimum frequency of golf ball numbers, maximum frequency of golf ball numbers, and the variance of frequency among golf ball numbers, in order.

```r
# Create a summary data
df <- matrix(c(golfMinFreq, golfMaxFreq, golfVar), nrow = NumberOfSims, ncol = 3)
colnames(df) <- c("min", "max", "variance")
rownames(df) <- 1:NumberOfSims
df <- as.table(df)
head(df)
```

```
##         min       max  variance
## 1 102.00000 136.00000 217.66667
## 2 109.00000 136.00000 196.33333
## 3 113.00000 138.00000 133.66667
## 4 115.00000 133.00000  68.33333
## 5 115.00000 129.00000  39.00000
## 6 107.00000 137.00000 153.00000
```

2. Observed Data

Second, all 3 test statistics are calculated for the observed data. The results are displayed below. The minimum frequency of golf ball numbers is **104**, the maximum frequency of golf ball numbers is **138**, and the variance of frequency among golf ball numbers is **343**.

```
# our dataset
ourData<- c(137, 138, 107, 104)

# Calculate observed data test statistics value
dataMinFreq <- min(ourData)
dataMaxFreq <- max(ourData)
dataVar <- var(ourData)

# Print as table
dfData <- matrix(c(dataMinFreq, dataMaxFreq, dataVar), nrow = 1, ncol=3)
colnames(dfData) <- c("min", "max", "variance")
rownames(dfData) <- 1
dfData <- as.table(dfData)
dfData
```

```
##   min max variance
## 1 104 138      343
```

3. p-value Calculation

p-values for each test statistics are calculated below. The p-value shows how likely are the observed golf ball counts, or to observe counts that are more extreme than the observed counts under the null hypothesis.

For the **minimum frequency of golf numbers**, we count the numbers that are smaller than the observed minimum frequency, **104**, and divide this by the total number of simulations, **10000**. Similarly, for the **maximum frequency of golf numbers**, we count the numbers that are larger than the observed maximum frequency, **138**, and divide this by the total number of simulations. Last, for the **variance of frequency among golf numbers**, we count the variances that are larger than the observed variance, **343**, and divide this by the total number of simulations.

```
# Calculated p-value by calculating the probability of encountering values more extreme than observed

# Sum overthe values that are more extreme than expected and average over total number of Simulations t
pMinFreq <- sum((golfMinFreq < dataMinFreq))/NumberOfSims
pMaxFreq <- sum((golfMaxFreq > dataMaxFreq))/NumberOfSims
pVar <- sum((golfVar > dataVar))/NumberOfSims
```

The calculated p-values for each statistic are displayed in the table below.

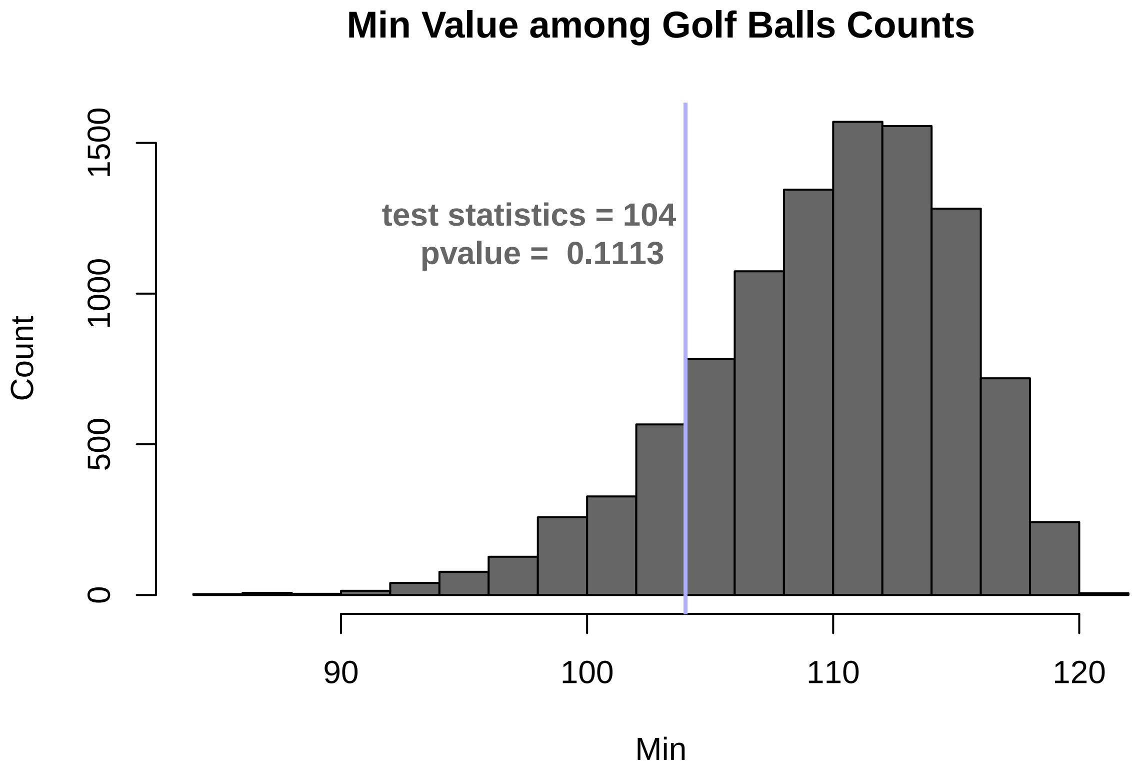
```
# Print as table
dfP <- matrix(c(pMinFreq, pMaxFreq, pVar), nrow = 1, ncol=3)
colnames(dfP) <- c("pval-min", "pval-max", "pval-variance")
rownames(dfP) <- 1
dfP <- as.table(dfP)
dfP
```

```
##   pval-min pval-max pval-variance
## 1   0.1113   0.1550        0.0373
```

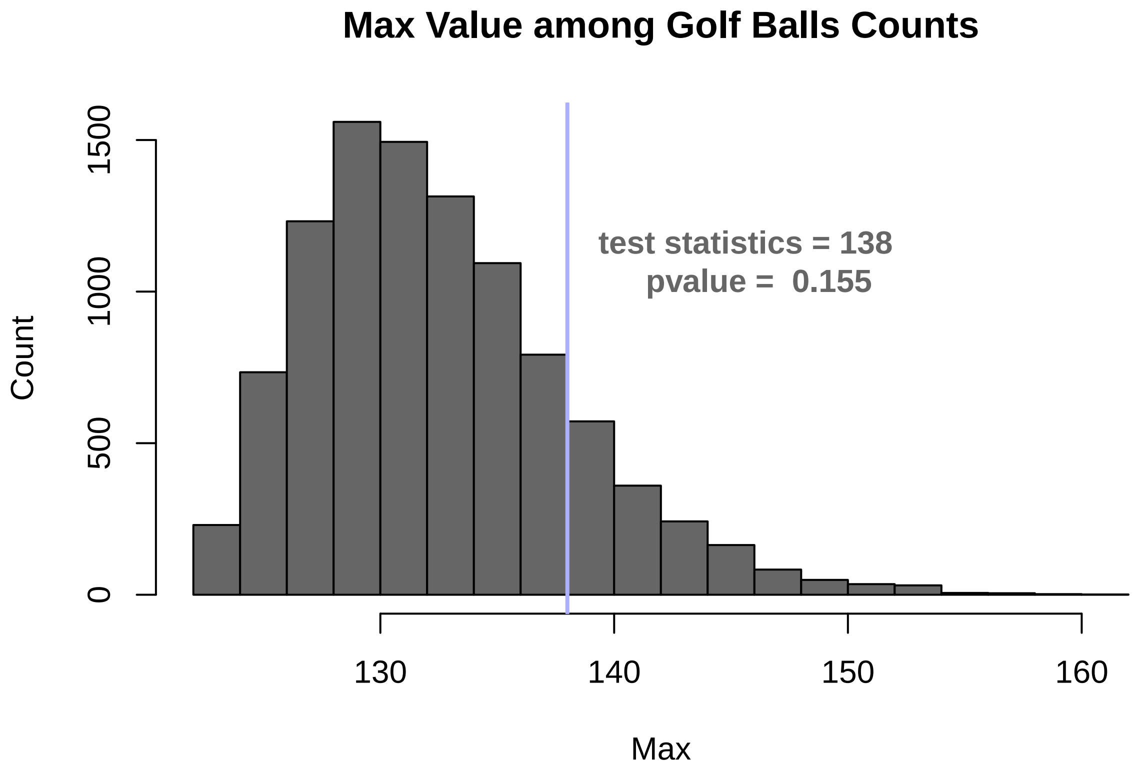
4. Histograms

Next, 3 histograms are created for each of the test statistics. The observed test statistic is plotted along with each histogram to show its value as well as its p-value.

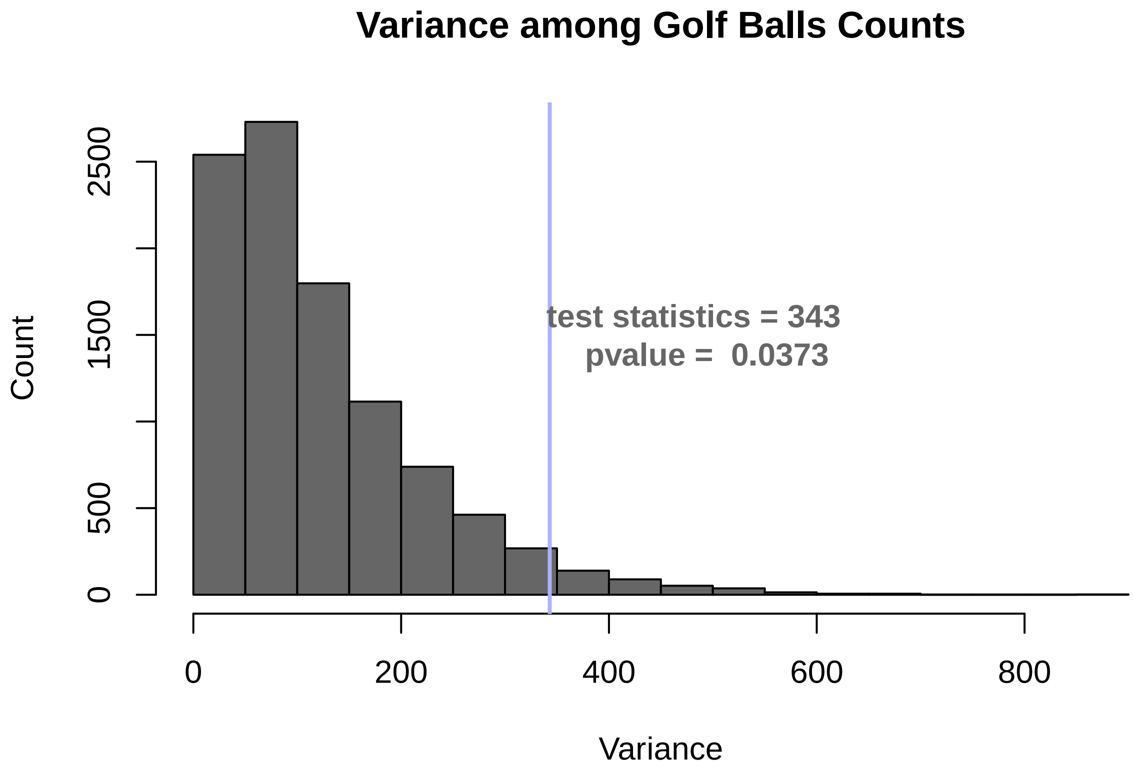
```r
# histogram plot highlighting the observed value
hist(golfMinFreq, main="Min Value among Golf Balls Counts", col = "grey40", xlab="Min", ylab="Count")
abline(v=dataMinFreq, col='#AFAFFF', lwd=2)
text(98,1200,paste("test statistics =",dataMinFreq," \n pvalue = ", pMinFreq), col="grey40",font=2)
```



```r
hist(golfMaxFreq, main="Max Value among Golf Balls Counts", col = "grey40", xlab="Max", ylab="Count")
abline(v=dataMaxFreq, col='#AFAFFF', lwd=2)
text(146,1100,paste("test statistics =",dataMaxFreq," \n pvalue = ", pMaxFreq), col="grey40",font=2)
```



```
hist(golfVar, main="Variance among Golf Balls Counts", col = "grey40",xlab="Variance", ylab="Count")
abline(v=dataVar, col='#AFAFFF', lwd=2)
text(490,1500,paste("test statistics =",dataVar," \n pvalue = ", pVar), col="grey40",font=2)
```



5. Chi-square test To assess the results of the chosen test statistics, the chi-square test is performed on the golf number counts. The observed counts are used as input to test against the null hypothesis. Since we assume a uniform distribution for the golf number count distribution under the null hypothesis, the probability of each golf number appearing is 0.25. The p-value for the test is calculated to be **0.03725**.

```
# Assume uniform distribution under null hypothesis, each number 1 through 4 have the same probability
nullprobs<- c(.25,.25,.25,.25)
(Xsq <- chisq.test(ourData, p=nullprobs))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  ourData
## X-squared = 8.4691, df = 3, p-value = 0.03725
```

```
# Prints test summary
Xsq
```

```
##
##  Chi-squared test for given probabilities
##
## data:  ourData
## X-squared = 8.4691, df = 3, p-value = 0.03725
```

III. Analysis

When the **minimum frequency** of golf balls among the 486 balls is used as the test statistics, 10000 simulations are performed and the 10000 simulated minimum frequency of golf balls are plotted into a histogram. The observed test statistics is calculated to be 104 and its p-value is calculated to be **0.1113**. Therefore, using a common p-value cutoff of 0.05, *we failed to reject the null hypothesis that the numbers 1, 2, 3, 4 are equally distributed.*

Similarly, when the **maximum frequency** of golf balls among the 486 balls is used as the test statistics, 10000 simulations are performed and the 10000 simulated maximum frequency of golf balls are plotted into a histogram. The observed test statistics is calculated to be 138 and its p-value is calculated to be **0.1550**. Therefore, using a common p-value cutoff of 0.05, *we failed to reject the null hypothesis that the numbers 1, 2, 3, 4 are equally distributed.*

Finally, when the **variance of frequency** among golf balls among the 486 balls is used as the test statistics, 10000 simulations are performed and the 10000 simulated variance of frequency among golf balls are plotted into a histogram. The observed test statistics is calculated to be 343 and its p-value is calculated to be **0.0373**. Therefore, using a common p-value cutoff of 0.05, we reject the null hypothesis that the numbers 1, 2, 3, 4 are equally distributed and *conclude that the numbers are not equally distributed.*

The test statistics **minimum frequency** and **maximum frequency** among golf balls both have a p-value greater than 0.05 so we failed to reject the null hypothesis. The test statistic **variance of frequency** among golf balls has a small p-value that allows us to reject the null hypothesis and accept the alternative hypothesis. Moreover, the p-value of this test statistic, **0.0373**, is very close to that calculated from the chi-square test, **0.03725**. One explanation is that the first two test statistics does not accurately describe the distribution of the numbers and the last test statistics has more power.

IV. Conclusion

In this project, we analyzed whether the golf ball numbers are evenly distributed given the counts of each number. 3 test statistics were chosen: the **minimum frequency** and **maximum frequency** of golf balls among the 486 balls, and the **variance of frequency** among golf balls. Then, 10000 rounds of simulations were run to obtain the p-value for each test statistics. Next, histograms were plotted to show the distribution of the test statistics from the simulation. The p-values for each test statistics were calculated. Last, these values were compared with the results of the theoretical test, the chi-squared test. The first 2 test statistics have large p-values so that we failed to reject the null hypothesis. However, the **variance of frequency** among golf balls test statistics achieved a small p-value that allows us to reject the null hypothesis. Moreover, it has a similar p-value to that of the chi-squared test. Thus, we reject the null hypothesis and conclude that the golf ball numbers are not evenly distributed.