Project #6

Ben Cartwright, Sally Kyong, Heewon Huh, Harvey Lee

2021-12-06

Problem #1 (20 points)

Independence of stock returns

Subsection 6.3.5 from your textbook looks at the following statement: *"Daily stock returns from the S&P500 for 10 days can be used to assess whether stock activity each day is independent of the stock's behavior on previous days"*. Your task is to re-do the work done in this section for a different index or stock. First, you would read and understand this section.

Next, you collect the data. One possibility is to look at a source like this one:

Yahoo Finance: Tesla

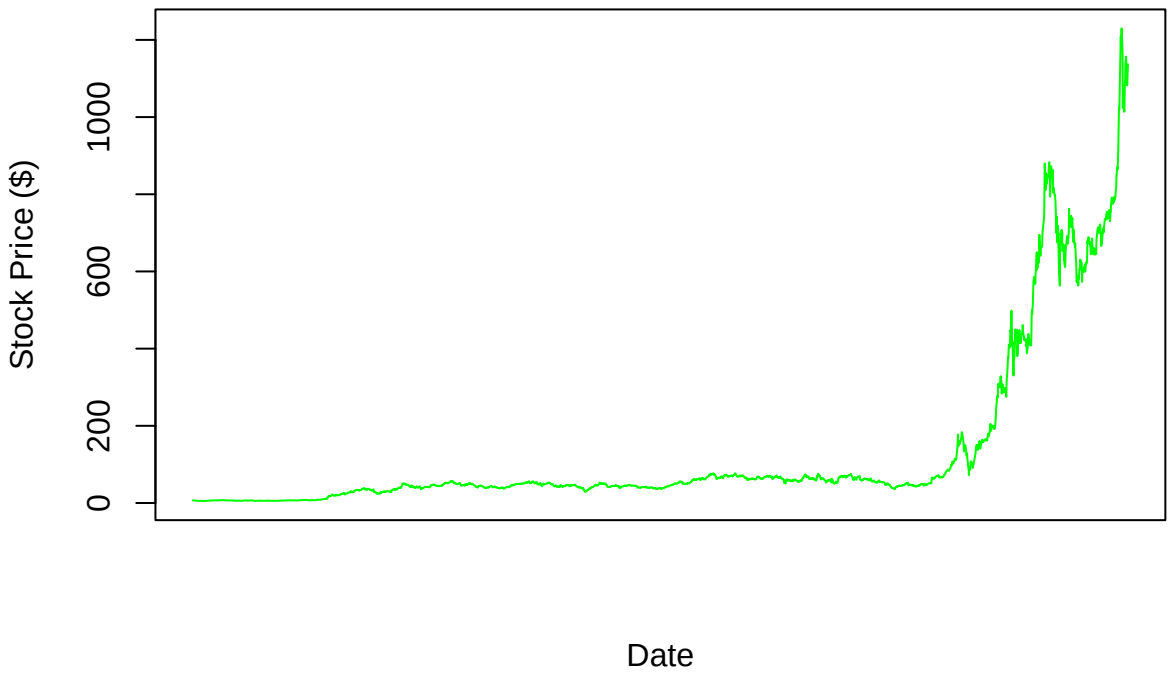
Then, you would download the data an create a nice time chart. It should look something like this:

Time plot

Then, you would mimic the analysis from Section 6.3.5 from the textbook and provide your conclusions.

Solution:

```
library(dplyr)
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
Tesla<- read.csv("TSLA.csv")

# Create the time plot for the daily movement of the stock prices over the past
# 10 years.
plot(as.Date(Tesla$Date), Tesla$Close, pch = 16, type = "l",
     main = "Daily Closing Stock Price of Tesla For Past 10 Years",
     xaxt = "n", xlab = "Date", ylab = "Stock Price ($)", col = "green")
```

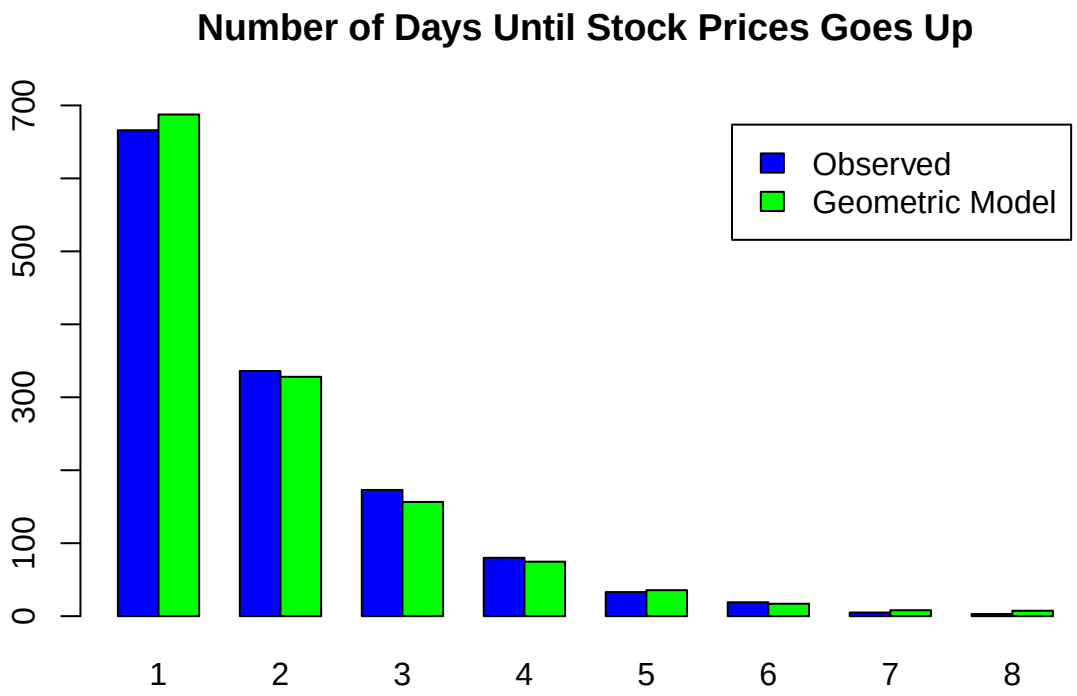



```
# Label the stock price movements as "Up" or "Down"
price<- Tesla$Close
change<- diff(price)
movement<-c()
for(i in 1:NROW(change)){
  if(change[i]<0){
    movement<-c(movement, "Down")
      }
  if(change[i]==0){
    movement<-c(movement, "No Change")
  }
  if(change[i]>0){
    movement<-c(movement, "Up")
  }
}


table(movement)
## movement
##      Down No Change        Up
##      1196         4      1315
# Count the number of days observed between each "Up" days.
count<- 1
days.to.up<- c()
for(i in 1:NROW(change)){
  if(change[i]<=0){
```

```
     count<- count+1
     days.to.up<- c(days.to.up,NA)
      }
  if(change[i]>0){
    days.to.up<-c(days.to.up, count)
    count<-1
  }
}


# Find the probability that the stock price goes up each day.
# Stock price went up 1315 days out of 2515 days.
n<- 1315
p<- 1315/2515


# Set up the geometric model for distribution of the number of days for the
# stock price to goes up.
p1<- p
p2<- (1-p)*p
p3<- (1-p)^(2)*p
p4<- (1-p)^(3)*p
p5<- (1-p)^(4)*p
p6<- (1-p)^(5)*p
p7<- (1-p)^(6)*p
p8<- 1-(p1+p2+p3+p4+p5+p6+p7)
geometric<- c(n*p1, n*p2, n*p3, n*p4, n*p5, n*p6, n*p7, n*p8)

t<- table(days.to.up)
t<- rbind(t,geometric)
rownames(t)<- c("Observed", "Geometric Model")
knitr::kable(t)
```

	1	2	3	4	5	6	7	8
Observed	666.0000	336.0000	173.0000	80.0000	33.0000	19.00000	5.000000	3.000000
Geometric Model	687.5646	328.0626	156.5309	74.6867	35.6358	17.00317	8.112843	7.403355

```
# Create the graph
barplot(t, beside = TRUE, main = "Number of Days Until Stock Prices Goes Up",
        legend = TRUE, ylim = c(0,700),
        col = c("blue", "green"))
```



H_0 : The stock market being up or down on a given day is independent from all other days. We will consider the number of days that pass until an Up day is observed. Under this hypothesis, the number of days until an Up day should follow a geometric distribution.

H_a: The stock market being up or down on a given day is not independent from all other days. Since we know the number of days until an Up day would follow a geometric distribution under the null, we look for deviations from the geometric distribution, which would support the alternative hypothesis.

The number of days to up has a multinomial distribution where each outcome is mutually exclusive from each other, so we will use the χ^2-test for goodness of fit to test our hypotheses.

```
# Do chi-square test for the goodness of fit.
a<- (666-687.5646)^2/687.4656
b<- (336-328.0626)^2/328.0626
c<- (173-156.5309)^2/156.5309
d<- (80-74.6867)^2/74.6867
e<- (33-35.6358)^2/35.6358
f<- (19-17.00317)^2/17.00317
g<- (5-8.112843)^2/8.112843
h<- (3-7.403355)^2/7.403355
q.sq<- a+b+c+d+e+f+g+h
1-pchisq(q.sq, 7)
## [1] 0.4061266
```

Since the $p-value = 0.4061266$ is large, there is not enough evidence against the null hypothesis, and we fail to reject the null hypothesis. The stock market being up or down on a given day is independent from all other days.

Problem #2 (10 points)

The Weight of Euro Coins

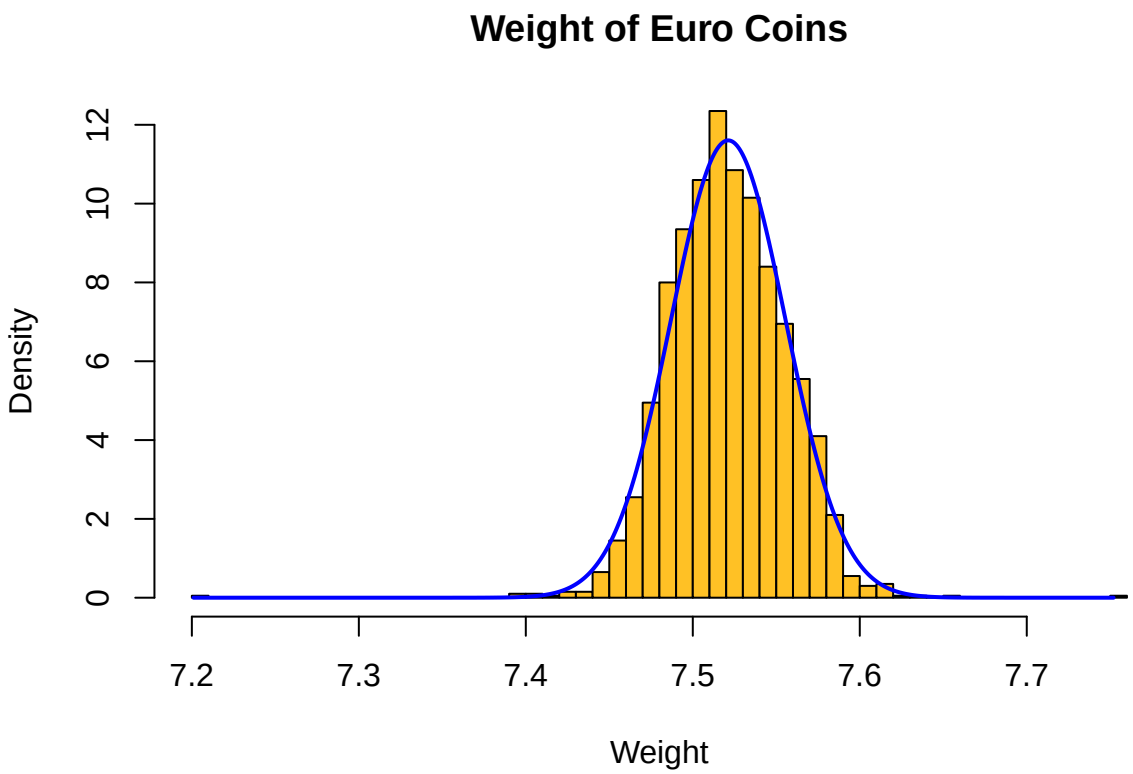
The paper *Ziv Shkedy, Marc Aerts & Herman Callaert (2006) The Weight of Euro Coins: Its Distribution Might Not Be As Normal As You Would Expect, Journal of Statistics Education, 14:2, DOI: 10.1080/10691898.2006.11910585* says **"According to information from the "National Bank of Belgium" the 1 Euro coin weighs 7.5 grams. It was anticipated that the weight of this coin would be normally distributed with mean 7.5 g."**

The paper's authors gathered the data on the weights of 2000 Euro coins. The data are available on the course website in the file called "euro-weights.csv". Read in the data set.

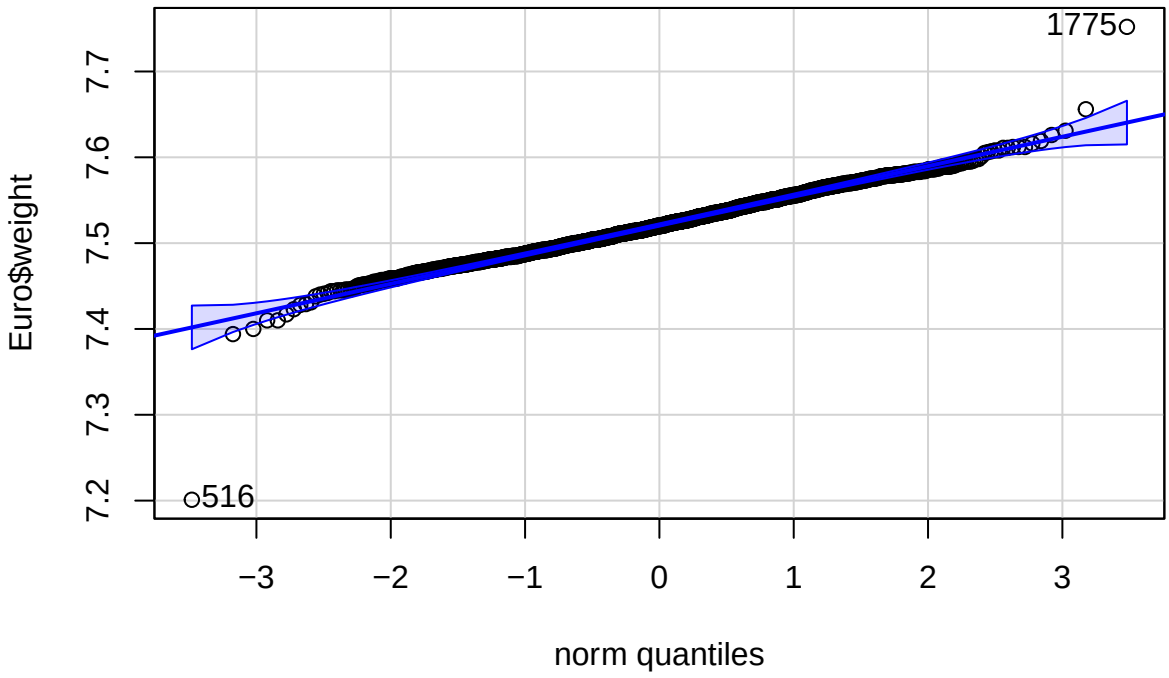
(2 points) Create an appropriate plot in `R` which will help you determine whether the distribution of the Euro-coin weights can be modelled as normal.

Solution:

```
Euro = read.csv("euro-weights.csv")

hist(Euro$weight, breaks = 50, prob = TRUE, main = "Weight of Euro Coins",
     xlab = "Weight", ylim=c(0,12), col = "goldenrod1")
x = seq(min(Euro$weight), max(Euro$weight), length = length(Euro$weight))
y = dnorm(x, mean = mean(Euro$weight),sd = sd(Euro$weight))
lines(x, y, col="blue", ylim=c(0,0.7), lwd = 2)
```



```r
install.packages("car", , repos = "http://cran.us.r-project.org")
## Installing package into 'C:/Users/heewo/OneDrive/Documents/R/win-library/4.1'
## (as 'lib' is unspecified)
## package 'car' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\heewo\AppData\Local\Temp\RtmpaMgGiE\downloaded_packages
library(car)
## Loading required package: carData
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
qqPlot(Euro$weight)
```




```
## [1]  516 1775
# qqPlot function shows that the most points fall approximately along the straight line,
# so we can assume normality.
```

(3 points) You would like to test the hypothesis that the mean Euro weight is 7.5 grams. Which test would you use? Justify why you would be able to use the test you propose. *Solution:*

If we were to test the hypothesis that the mean Euro weight is 7.5 grams, then we probably would use the t-test because of the following reasons: 1. We don't know the true population standard deviation. 2. The sample size is greater than 30. 3. The sample is approximately normal.

Since all the requirements are met for t-test, we should perform the hypothesis test using t-test.

(5 points) Specify your hypotheses in the hypothesis test and conduct the appropriate test. Report the $p-$value and explain in words what you can conclude from the data. *Solution:*

$H_0 : \mu = 7.5$ $H_A : \mu \neq 7.5$

```
# Find the sample mean and standard deviation
x.bar = mean(Euro$weight)
sd = sd(Euro$weight)

# Find the degrees of freedom and t-statistics
df = 2000 - 1
t = (x.bar - 7.5)/(sd/sqrt(2000))

# Find the p-value
p_value = 2 * (1 - pt(t, df))
print(p_value)
## [1] 0
# Since the p-value is really small, there is strong evidence against the null hypothesis, hence we rej
```

Problem #3 (15 points)

Case study: Malaria vaccine

Section 2.3 in our textbook contains a case study about the efficacy of a malaria vaccine. Again, since you know the appropriate tests, you can now conduct the analysis.

The original data set is available at

Malaria data

(3 points) First, provide a paragraph or two about the historical background.

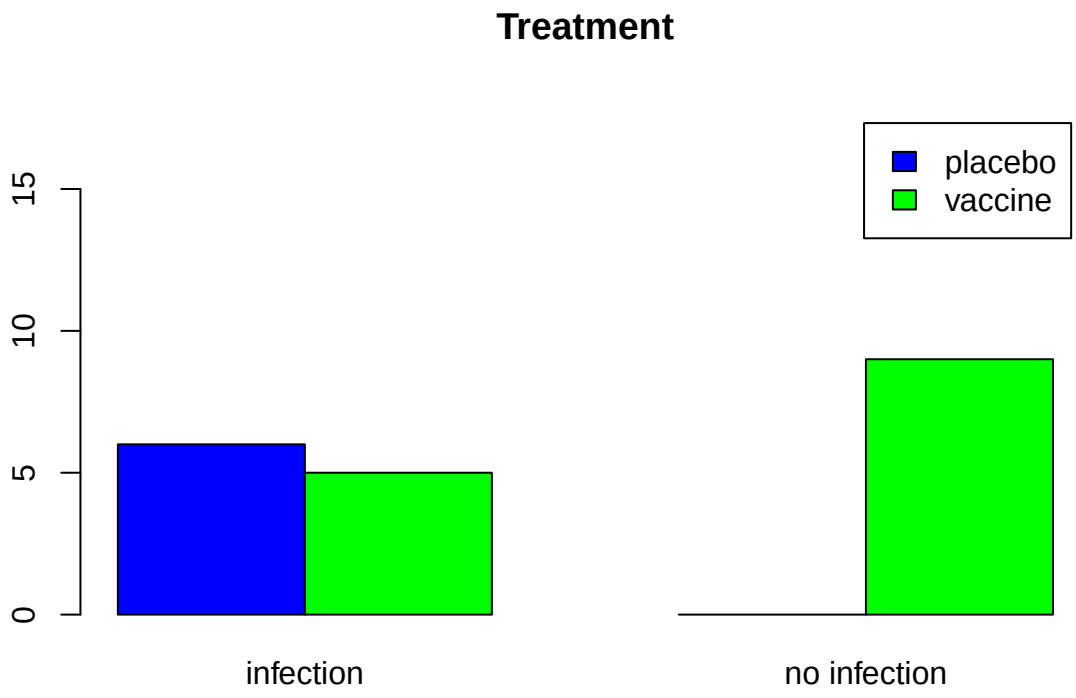
The malaria disease came from the mosquitoes that carried the malaria parasite, which flew to humans for thousands of years. As half of the world's population lives in areas vulnerable to malaria, there have been many different attempts to defeat the disease. One of them was quinine, a substance from the cinchona tree, effective since the 1600s. After some time, the scientists started to focus on vector control as they had a better understanding of malaria transmission. Also, there have been different antimalarial drugs to fight against the disease. By taking these measures, the data collected by the scientists showed that the deaths from malaria fell 13% from 2000 to 2010. Although there are different ways to treat malaria, it remains a problem because of the emergence of resistance to drugs and insecticides. Therefore, scientists started to focus on malaria vaccines to decrease the number of malaria cases and death rates.

(3 points) Then, visualize the data (differently from the textbook figures, please).

Solution:

```
data = read.csv("malaria.csv")
t = table(data$treatment, data$outcome)
print(t)
##
##           infection no infection
##   placebo         6            0
```

```
##   vaccine         5            9
barplot(t, beside = TRUE, main = "Treatment", legend = TRUE, ylim = c(0,18), col = c("blue", "green"))
```



(3 points) Then, follow with a research question and proposed statistical procedure.

Solution:

Research Question: To see whether the vaccine is effective or not, we want to test whether the treatment and outcome are independent.

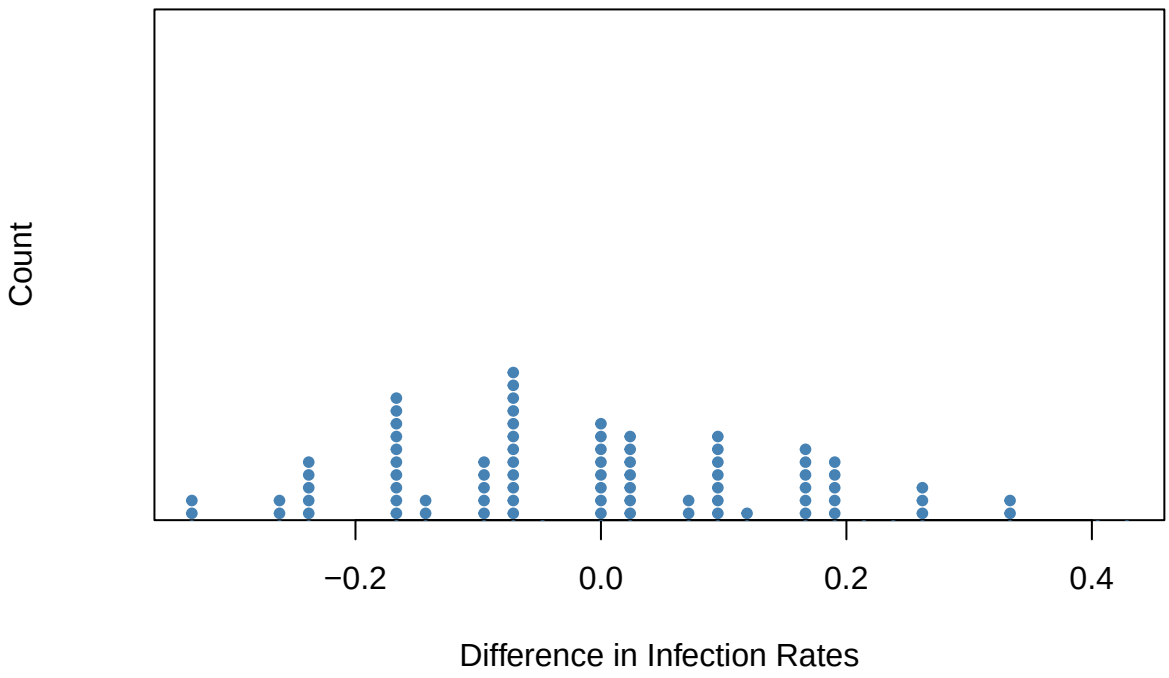
H_0 :: **Independence model**. The variables treatment and outcome are independent. They have no relationship, and the observed difference between the proportion of patients who developed an infection in the two groups, 64.3%, was due to chance.

H_a :: **Alternative model**. The variables are not independent. The difference in infection rates of 64.3% was not due to chance, and vaccine affected the rate of infection.

We are going to simulate the study assuming the malaria vaccine does not work.

```
sim<- function(){
 # Create the sample of 20 patients.
 # Let 1-11 represents "infection" and 12-20 represents "no infection".
 patients<- c(1:20)

 # Randomly select 14 vaccinated and 6 placebo patients.
 vaccinated<- sample(patients, 14)
```

```
 placebo<- patients[-vaccinated]

 # Count the number of "infection" and "no infection" in each group.
 v.inf<-0
 v.noinf<-0
 for(i in 1:length(vaccinated)){
   if(vaccinated[i]<=11){
     v.inf<- v.inf+1
   }
   if(vaccinated[i]>11){
     v.noinf<- v.noinf+1
   }
 }
 v.outcome<-c(v.inf, v.noinf)

 p.inf<-0
 p.noinf<-0
 for(i in 1:length(placebo)){
   if(vaccinated[i]<=11){
     p.inf<- p.inf+1
   }
   if(vaccinated[i]>11){
     p.noinf<- p.noinf+1
   }
 }
 p.outcome<-c(p.inf, p.noinf)

 # Find the difference in infection rate
 diff<- (p.inf/length(placebo))-(v.inf/length(vaccinated))
}

# Run 100 simulations.
run<-replicate(100, sim())
stripchart(run, at=0, pch=20,
           method="stack", col="steelblue",
           xlab="Difference in Infection Rates",
           ylab="Count",
           ylim = c(1,20))
```



In a given data set, the difference in infection rates was:

$\frac{6}{6} - \frac{5}{14} = 0.642857143$

Under the assumption that vaccine does not work, our replication shows that the difference of 0.642857143 is a rare event. Since the difference in infection rate of 0.642857143 is unlikely to be just a random occurrence by chance, we concluded that the vaccine does work and the treatment and outcomes are dependent, hence reject the null hypothesis.

(9 points) *Caveat:* You cannot just use the test we used for similar research questions **Why?**. You must develop your own test using the same logic we used to construct the usual test. Would the multinomial distribution help? What are your conclusions?

Solution:

The simulation model can be expensive and take a significant amount of computer time as the sample size increases. Instead, we can use Fisher's exact test. Knowing the 11 of 20 patients are infected and 14 of 20 are vaccinated and assuming the vaccine does not work, we want to know the probability that the proportion of "infection" in two groups are uneven. Use the same hypothesis from the previous section:

H_0 :: **Independence model**. The variables treatment and outcome are independent. They have no relationship, and the observed difference between the proportion of patients who developed an infection in the two groups, 64.3%, was due to chance.

H_a :: **Alternative model**. The variables are not independent. The difference in infection rates of 64.3% was not due to chance, and vaccine affected the rate of infection

```
malaria<-read.csv("malaria.csv")
t = table(malaria$treatment,
          malaria$outcome)
colnames(t)<- c("Infection", "No Infection")
rownames(t)<- c("Placebo", "Vaccinated")
knitr::kable(t)
```

	Infection	No Infection
Placebo	6	0
Vaccinated	5	9

Using the hypergeometric distribution, the probability that we see the given outcome is:

$p = \frac{\binom{6}{6}\binom{14}{5}}{\binom{20}{11}}$

```
(length(combn(6,6))*length(combn(14,5)))/length(combn(20,11))
## [1] 0.03250774
```

Since the given out come has the low probability of occurrence of 0.03250774, we think it is unlikely to happen purely by the random chance. Therefore, we think our initial assumption is incorrect. The vaccine does work and we reject the null hypothesis; the treatment and outcome are dependent. The use of multinomial distribution like χ^2-test fives the approximation of significance value, but the approximation is inadequate when the sample size is small like the given study. Therefore, the multinomial distribution does not help in this case.

Problem #4 (15 points)

Physicians' Reactions to Patient Size

The citations of the original paper we are interested in are available at

Hebl-Xu: *Weighing the care: physicians' reactions to the size of a patient*

For more information about the experiment performed in the paper, please, read the background available from the

Rice Virtual Lab in Statistics.

The background is linked here:

Background

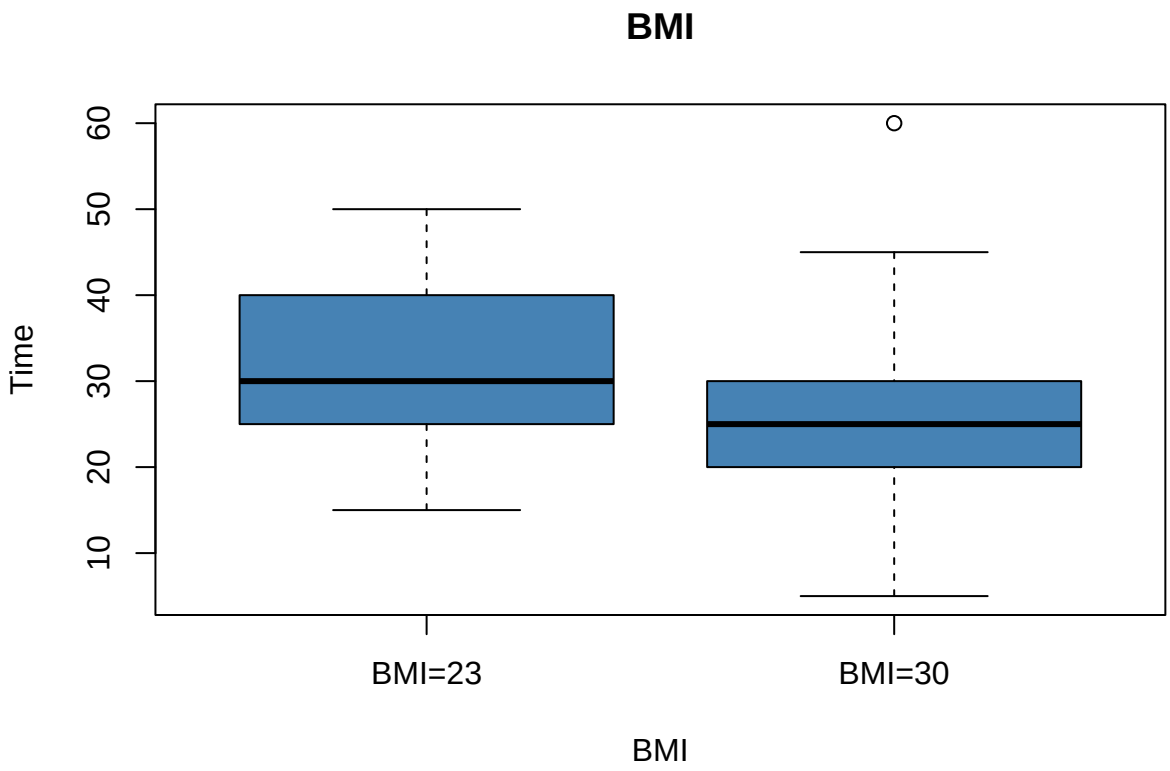
The data associated with the above paper are available at

Patient data

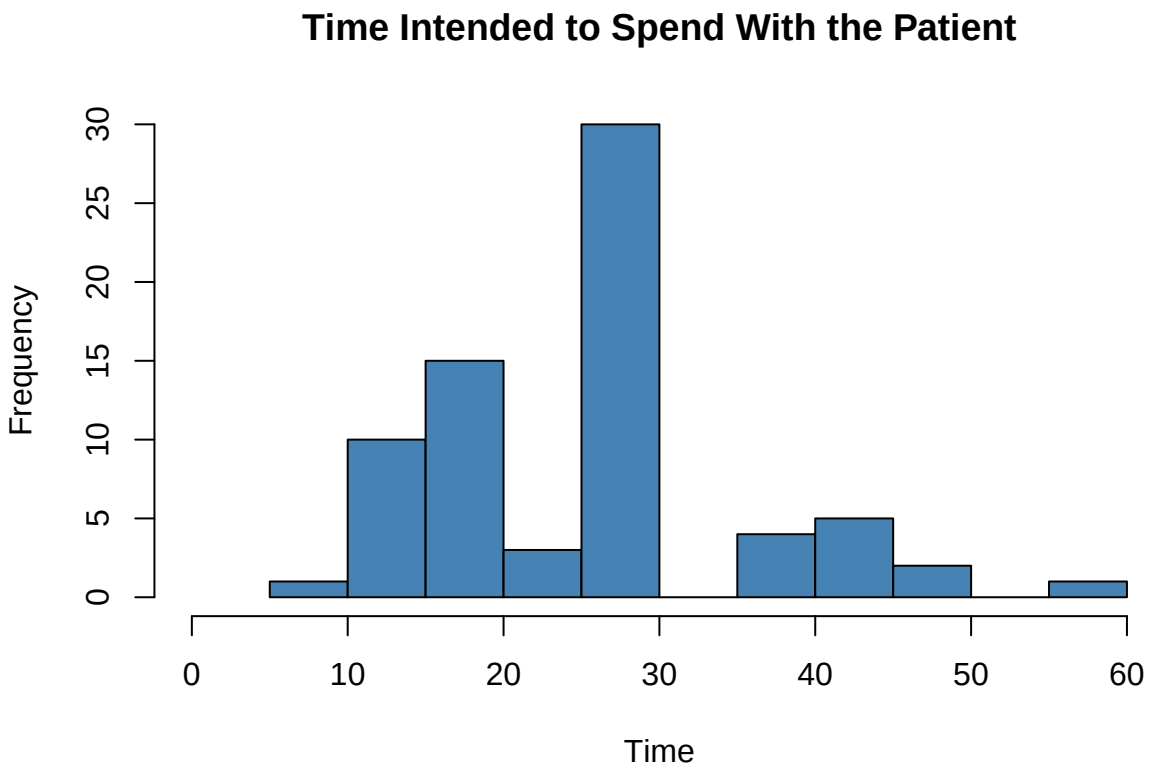
(4 points) Read in the data and create side-by-side boxplots of the time intended to spend with the patient for the two groups of patients: the ones with the lower BMI and the ones with the higher BMI. Plot the histograms of the time intended to spend with the patient for the two groups of patients for the two groups of patients as well. What features do the box plots and the histograms have?

Solution:

```r
BMI = read.csv("discriminate.csv")

boxplot(BMI$time ~ BMI$weight, col = "steelblue", main = "BMI", xlab = "BMI", ylab = "Time")
```




```r
hist(BMI$time, main = "Time Intended to Spend With the Patient", xlab = "Time", col = "steelblue", xlim
```



One main feature of boxplot is that it shows different quartiles, as well as showing outliers which is helpful in identifying the skewness of the data. Whereas, the main feature of the histogram displays the frequencies of a data set, that is, it is useful when there is a variance difference.

(4 points) Your goal is to see if there is an effect of the weight of the patient on the mean time the physician intends to spend with the patient. Formulate your hypotheses based in this research question.

Solution:

```
mean(BMI$time)
## [1] 27.8169
```

$H_0 : \mu \leq 27.8169$ $H_a : \mu > 27.8169$

(3 points) Which statistical procedure do you plan to use? Justify why you are allowed to use the said procedure.

Solution:

The best statistical procedure for this particular research would be using the t-test for the following reasons: 1. T-test is really useful when comparing the means of the two groups; in this particular research, it would be comparing the means of 2 different BMI groups to see the mean time the physician intends to spend with the patient upon their BMI group. 2. The sample is random. 3. The population size is bigger than 30, so we assume the sample is approximately normal.

(4 points) Perform the appropriate test on your data, report the $p-$value, and summarize your findings.

Solution:

```
t.test(BMI$time)
##
##  One Sample t-test
##
## data:  BMI$time
## t = 22.895, df = 70, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  25.39370 30.24011
## sample estimates:
## mean of x
##   27.8169
```

Since the p-value is really small there is strong evidence against the null hypothesis. Therefore, we should reject the null hypothesis.