Project #6

Antoinette Lynne, Harini Shanmugam, Aanya Shrivastava

2021-12-07

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

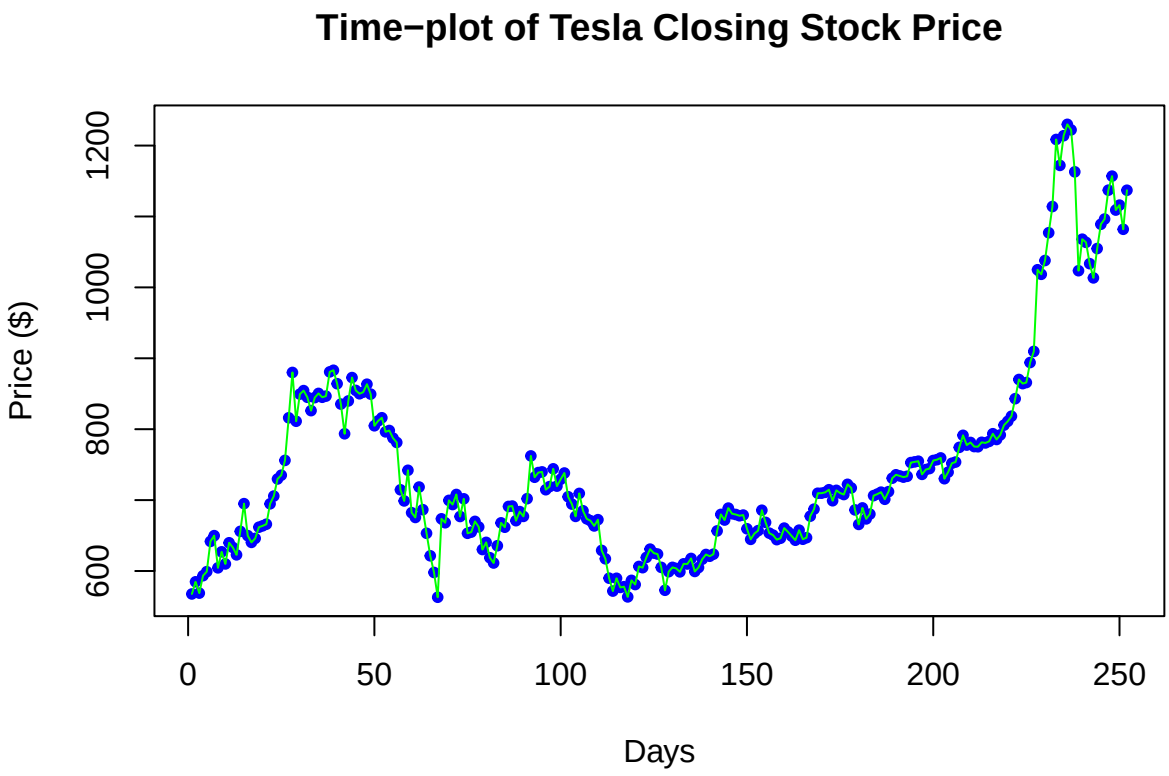
Then, you would download the data an create a nice time chart. It should look something like this:

Time plot

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

```
# time chart
tesladata = read.csv("TSLA.csv")

closingprice = tesladata$Close
plot(closingprice,
     main="Time-plot of Tesla Closing Stock Price",
     xlab="Days",
     ylab="Price ($)",
     pch=20,
     col="blue")
lines(closingprice, col="green")
```




```
# analysis mimicked from Section 6.3.5
closingprice = tesladata$Close
openingprice = tesladata$Open

n = length(closingprice)
pricechange = c()
directionchange =c()

for(i in 1:n) {
  pricechange.initial = closingprice[i] - openingprice[i]
  pricechange = append(pricechange.initial, pricechange)
  if(pricechange.initial > 0) {
    directionchange = append("Up", directionchange)
  } else {
    directionchange = append("D", directionchange)
  }
}

days = function(index, numofdaysvector, directionchange) {
  numofdays = 1
  index = index + 1
  while(identical(directionchange[index], "D")) {
    numofdays = numofdays + 1
    index = index + 1
  }
  return(numofdays)
```

```
}

numofdaysvector = c()
for(i in 1:n) {
    numofdays = days(i, numofdaysvector, directionchange)
    numofdaysvector = append(numofdays, numofdaysvector)
}

# $H_0$: 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.
# vs
# $H_a$: 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.

expectedvalue = c()
for(i in 1:7) {
  pvalue = ((1-0.545)^(i-1)) * (0.545) * i
  expectedvalue = append(ceiling(pvalue), expectedvalue)
}
expectedvalue = rev(expectedvalue)
expectedvalue
## [1] 1 1 1 1 1 1 1
library(plyr)
observedvaluetable = count(numofdaysvector)
observedvaluetable
##   x freq
## 1 1  128
## 2 2   67
## 3 3   32
## 4 4   13
## 5 5    9
## 6 6    2
## 7 7    1
observedvalue = c(128, 67, 32, 13, 19, 2, 1)

chisquare = 0
for(i in 1:7) {
  chisquare = chisquare + ((observedvalue[i] - expectedvalue[i])^2) / expectedvalue[i]
}

significance.10 = qchisq(0.1, 6, lower.tail = FALSE)
significance.10
## [1] 10.64464
significance.5 = qchisq(0.05, 6, lower.tail = FALSE)
significance.5
## [1] 12.59159
significance.1 = qchisq(0.01, 6, lower.tail = FALSE)
significance.1
## [1] 16.81189
```

```
# For all three significance levels, the calculated chi square values are
# greater than the expected values. Thus, we reject the null hypothesis.
# Meaning, the number of days that pass until an "up" day does not follow a
# geometric distribution and the event of the stock market being an "up" or
# "down" day is not independent.
```

Problem #2 (10 points)

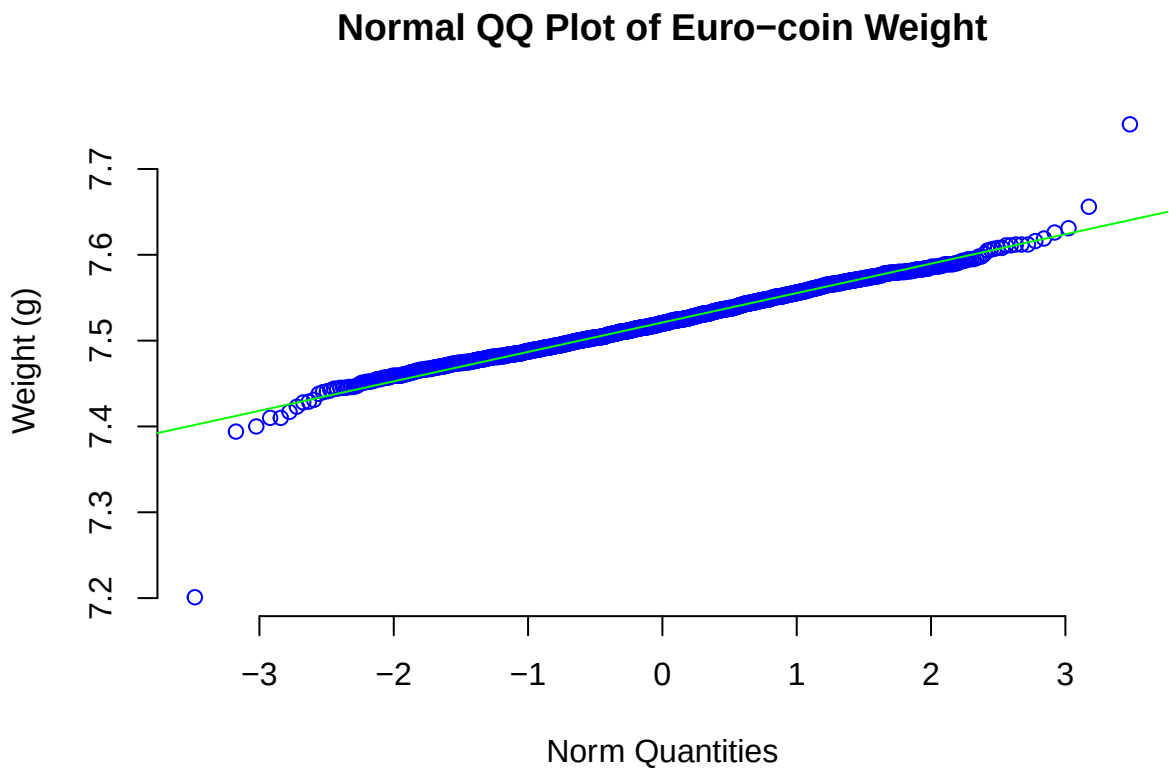
The Weight of Euro Coins

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.

```
eurodata = read.csv("euro-weights.csv")
weight <- eurodata$weight[eurodata$index]
qqnorm(weight,
       main="Normal QQ Plot of Euro-coin Weight",
       xlab="Norm Quantities",
       ylab="Weight (g)",
       pch=1,
       col="blue",
       frame=FALSE)
qqline(weight,
       col= "green")
```



(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.
We would use a z-test since the mean and standard deviation are known and the sample is very large.

(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. $H_0 : \mu = 7.5$ vs $H_a : \mu \neq 7.5$

```
z_stat = (mean(weight)-7.5) / (sd(weight))
pnorm(z_stat)
## [1] 0.7315799
```

Since the p-value is greater than 0.05, we fail to reject the null hypothesis. Meaning, the mean Euro coin weight is indeed 7.5.

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

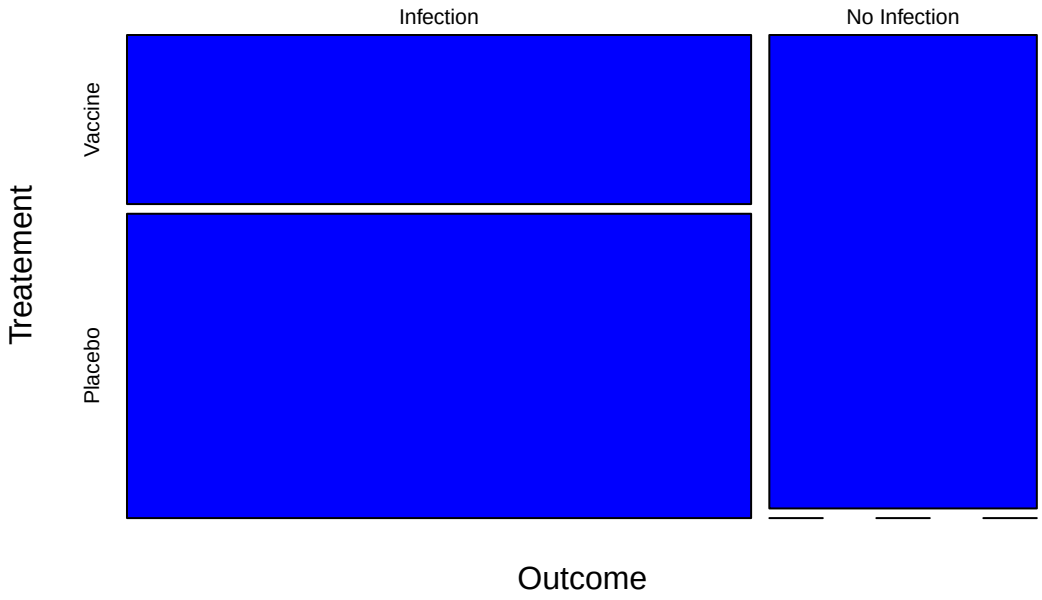
(3 points) First, provide a paragraph or two about the historical background.
Malaria is a mosquito borne infectious disease caused by Plasmodium parasites. More specifically, it is spread by the saliva of female Anopheles mosquitoes carrying the parasite. The side effects of the disease include fever, chills, and flu-like symptoms. Left untreated, malaria can lead to death. However, if treated promptly with antimalarial drugs, the disease can be cured in two weeks.
In 2019, there were about 229 million cases of malaria globally. 409,000 of these cases resulted in death, 274,000 of which were children under the age of 5 who were mostly in sub-Saharan Africa. The disease is most frequently transmitted in tropical and subtropical areas. Although the United States does not have any tropical or subtropical regions, the disease is still transmitted through residents travelling to and from countries near the equator. Only in October 2021 did the World Health Organization approve of a vaccine against malaria for the first time. The plan is to vaccinate people and mainly children in sub-Saharan Africa and other high transmission regions first. So only time will tell what the course for malaria will look like in the 21st century.

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

```
malaria = matrix(c(5,6,9,0), ncol=2)

rownames(malaria) = c('Infection','No Infection')
colnames(malaria) = c('Vaccine', 'Placebo')

mosaicplot(malaria,
           main="Summary Results for Malaria Vaccine Experiment",
           xlab="Outcome",
           ylab="Treatement",
           col="blue")
```



(3 points) Then, follow with a research question and proposed statistical procedure.
Research question: Is this malaria vaccine effective? Does it work or not? Proposed statistical procedure: We must test if the treatment and outcome are independent.
H_0: the treatment and outcome are independent vs H_a: the treatment and outcome are not independent

(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?
We can not use a chi-square test because it does not pass the necessary initial conditions. Straight away, we can not use the chi-square test because the sample size (20) is too small. In this situation, since the sample size is relatively small, we would use the Fisher's exact test. We can use it because it passes the condition that the row and column totals in the two way table are fixed and that at least one of the values in the two way table is less than 5. The multinomial distribution could help since it would show if the results of the placebo treatment and vaccine treatment are normal on their own.

```
malaria = matrix(c(5,6,9,0), ncol = 2)

rownames(malaria) = c('Infection','No Infection')
colnames(malaria) = c('Vaccine', 'Placebo')

malariatable = as.table(malaria)
malariatable
##              Vaccine Placebo
## Infection          5       9
## No Infection       6       0
fisher.test(malariatable)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  malariatable
## p-value = 0.01409
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0000000 0.7471714
## sample estimates:
## odds ratio
##          0
```

Since the calculated p-value is less than 0.05, we reject the null hypothesis. Meaning that the treatment and outcome are not independent.

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

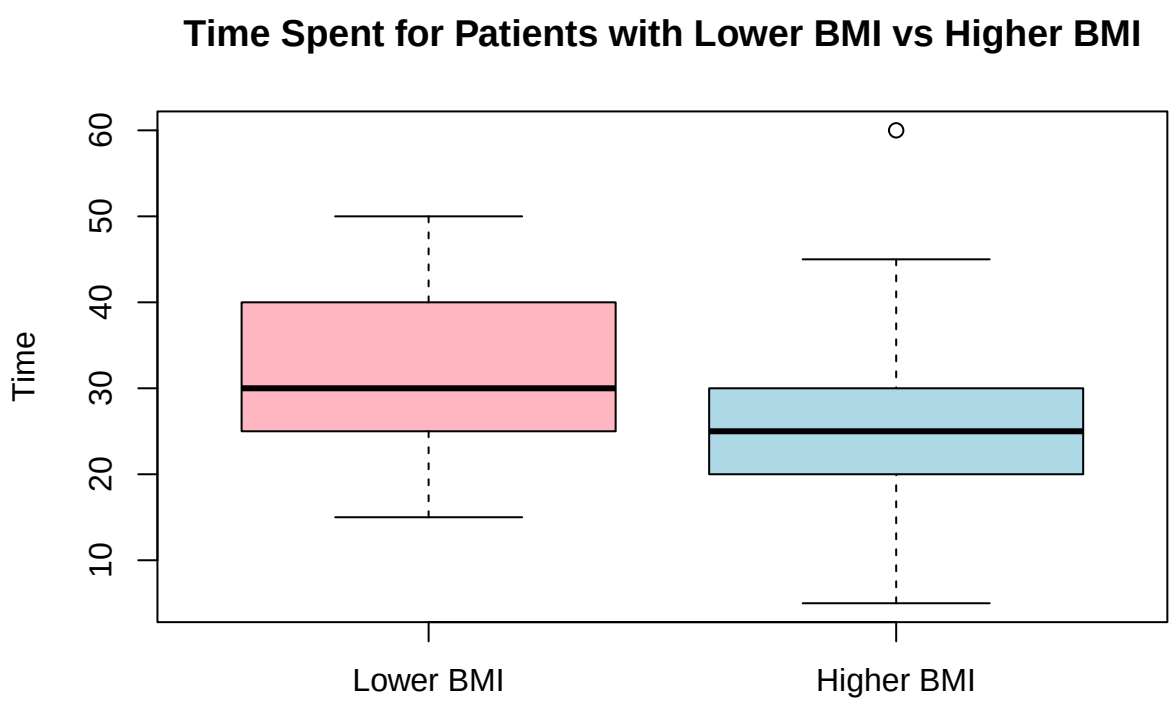
The data associated with the above paper are available at

Patient data

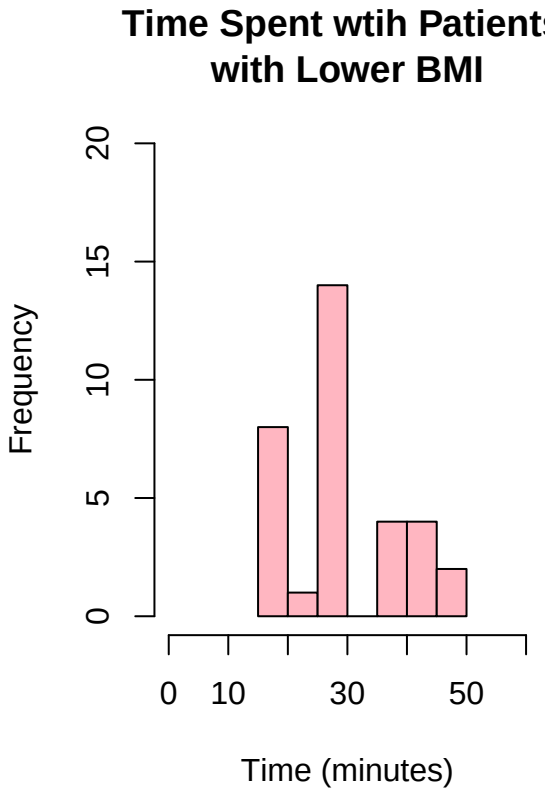
(4 points) Read in the data and create side-by-side boxplots of the time intended to spend with the patient for the two groups of patients: the ones with the lower BMI and the ones with the higher BMI. Plot the histograms of the time intended to spend with the patient for the two groups of patients for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

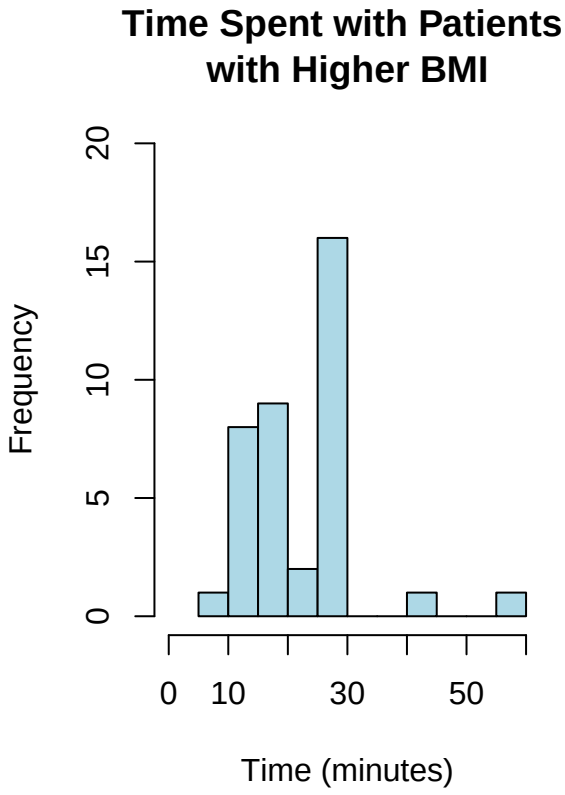
```
patientdata = read.csv("discriminate.csv")
lowerbmi = patientdata$time[patientdata$weight == "BMI=23"]
higherbmi = patientdata$time[patientdata$weight == "BMI=30"]

par(mfrow=c(1,1))
boxplot(lowerbmi, higherbmi,
        main="Time Spent for Patients with Lower BMI vs Higher BMI",
        ylab="Time",
        names=c("Lower BMI", "Higher BMI"),
        col=c("lightpink", "lightblue"))
```




```
par(mfrow=c(1,2))
hist(lowerbmi,
          main="Time Spent wtih Patients \nwith Lower BMI",
          xlab="Time (minutes)",
          ylab="Frequency",
          xlim=c(0,60),
          ylim=c(0,20),
          breaks=12,
          col="lightpink")
hist(higherbmi,
          main="Time Spent with Patients \nwith Higher BMI",
          xlab="Time (minutes)",
          ylab="Frequency",
          xlim=c(0,60),
          ylim=c(0,20),
          breaks=12,
          col="lightblue")
```





Both pairs of boxplots and histograms are slightly skewed right.
In the pink box plot representing the lower BMI, the range between the 25th percentile and the 50th percentile is much smaller than the other half of the interquartile range. This shows that there are a lot more data points within this small range, making the distribution skewed to the right. Similarly, in the blue box plot representing the higher BMI, there is an outlier at around the 60 minutes mark which skews the distribution to the right.
The histograms correspond to their boxplot counterparts are also skewed to the right. Additionally, the mode for both histograms is the interval of 25-30 minutes.

(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.
$H_0 : \mu_{23} = \mu_{30}$ vs $H_a : \mu_{23} > \mu_{30}$

(3 points) Which statistical procedure do you plan to use? Justify why you are allowed to use the said procedure.
We would use an unpaired two sample t-test because we are seeing if the means of the two groups are not equal, which would indicate a difference in the amount of time a physician spends with their patient depending on their BMI. In this case, we are testing to see if the mean of time spent with patients with a lower BMI is higher than time spent with patients with a higher BMI.

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

```
mu_23 = mean(lowerbmi)
mu_23
## [1] 31.36364
mu_30 = mean(higherbmi)
mu_30
```

```
## [1] 24.73684
t.test(lowerbmi, higherbmi, var.equal = TRUE)
##
##  Two Sample t-test
##
## data:  lowerbmi and higherbmi
## t = 2.856, df = 69, p-value = 0.005663
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.997955 11.255633
## sample estimates:
## mean of x mean of y
##  31.36364  24.73684
```

Since the calculated p-value is less than 0.05, we reject the null hypothesis. Meaning that μ_{23} is greater than μ_{30}.