Stat 21 HW 1

Dohyun Lee

3. Blood Pressure (12 Points)

Many years ago, I worked on a study on women with polycystic ovarian syndrome (PCOS), an endocrine disease related to diabetes (Legro, Bentley-Lewis, Driscoll, Wang, and Dunaif (2002): J. Clinical Endocrinology and Metabolism 87:5). As part of the study, 371 women with PCOS were selected. Do women with this condition tend to differ in blood pressure from the general population? The 371 women had a mean systolic blood pressure of 121.07, with an SD of 16.59. Let μ denote the mean blood pressure of the population from which these 371 subjects were selected. (You may assume that these subjects can be considered to be representative of some larger population.)

(a) A "normal" systolic blood pressure is considered to be 120. Carry out a hypothesis test of the following hypotheses using an alpha level of .05: H_0: $\mu = 120$ vs. H_A: $\mu \neq 120$ Be sure to label the test statistic and the p-value. Do these control subjects significantly differ from a "normal" population in systolic blood pressure?

1)

H_0: $\mu = 120$

H_A: $\mu \neq 120$

2)

Test Statistic:

$\bar{X} = 121.07$

$\mu = 120$

$s = 16.59$

t' $= \frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}}$

t' $= \frac{121.07-120}{\sqrt{\frac{16.59^2}{371}}}$

t' = 1.244

p-value:

```r
2*pt(-1.244, df = 370)
```

```
## [1] 0.2142871
```

The p-value is 0.214 which is more than 0.05. So, we fail to reject H_0 and we can say that there is no significant difference between the mean systolic blood pressure of the control subjects and the "normal" systolic blood pressure.

(b) If the sample size had been n = 3710 instead of 371, how would your conclusions change? How does this illustrate the concept of statistical significance as distinct from effect size — that is, the size of the effect in a medical setting?

t' = $\frac{121.07-120}{\sqrt{\frac{16.59^2}{3710}}}$ t' = 3.928

```r
2*pt(-3.928, df = 3709)
```

```
## [1] 8.721389e-05
```

We get a p-value of 0.00008 which is less than 0.05, which tells us to reject H_0. Effect size is the magnitude of the difference of the means. Statistical significance is how confident we are that something isn't occurring due to random chance. This tells us that, by keeping our effect size constant, having a larger sample size increases our statistical significance.

(c) Calculate a 95% confidence interval for μ. Does this interval contain 120? Is your confidence interval consistent with your conclusion in part (a)?

```r
SE <- 16.59/sqrt(371) #Calculate the standard error
t <- qt(0.025, df = 370) #Find our confidence coefficient
121.07 + c(t,-t)*SE #Calculate our confidence interval
```

```
## [1] 119.3763 122.7637
```

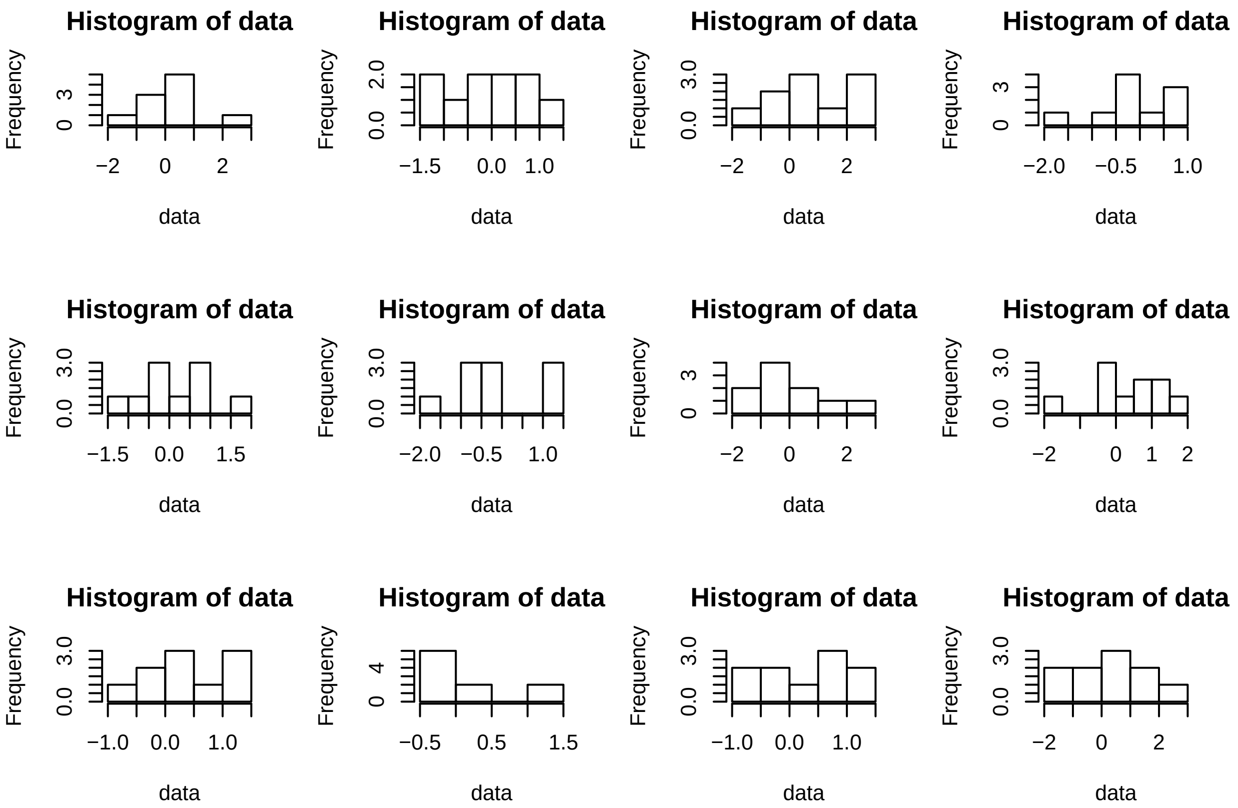
The confidence interval for μ is (119.38, 122.76) and this interval is consistent with our conclusion in part a because 120 in within our range. We are 95% confident that the average systolic blood pressure for women falls between 119.38 and 122.76.

4. Normal Probability Plots (10 Points)

(a) Use R to generate a random sample having 10 observations from a standard Normal distribution. Then make a histogram of the dataset. You can use the following commands:

Repeat this a total of twelve times. To plot all twelve histograms on one page, give this command before creating the histograms:

```r
par(mfrow=c(3,4)) # plot 12 plots per page in a 4-by-3 array in row order
for (i in 1:12) {
  data <- rnorm(10, 0,1) # generate sample from Normal population with n = 10, mean = 0, SD = 1
  hist(data) # make histogram
}
```

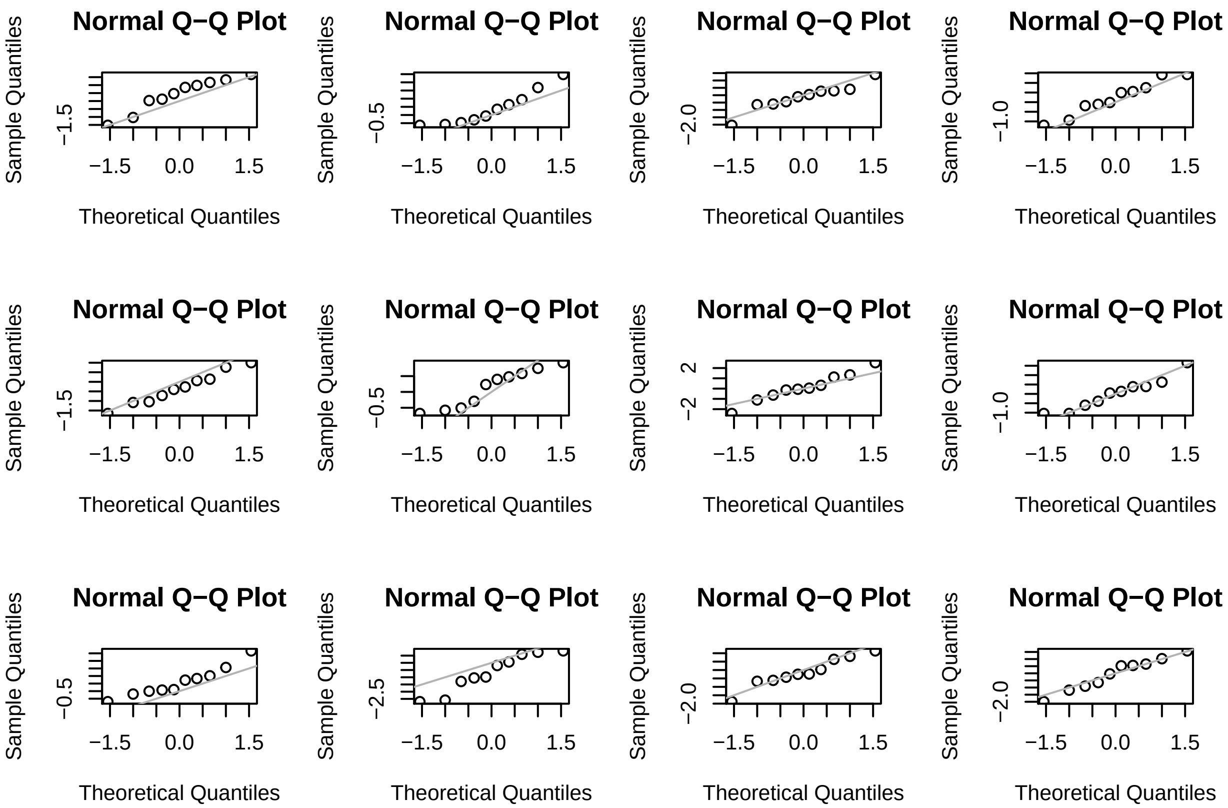


Do the samples appear to be normally distributed? Explain briefly. (Hand in the page with the twelve histograms.)

Not every sample appears to be normally distributed. Some appear to be somewhat normal and some are skewed.

(b) Now repeat the above using normal probability plots (NPP). You don't need to plot the same twelve datasets; just create twelve new datasets as above, and use these command to make the NPP:

```r
par(mfrow=c(3,4)) # plot 12 plots per page in a 4-by-3 array in row order
for (i in 1:12) {
  data <- rnorm(10, 0,1) # generate sample from Normal population with n = 10, mean = 0, SD = 1
  qqnorm(data) # make NPP
  abline(0,1, col=gray(.7)) # add gray line with intercept = 0, slope = 1
}
```

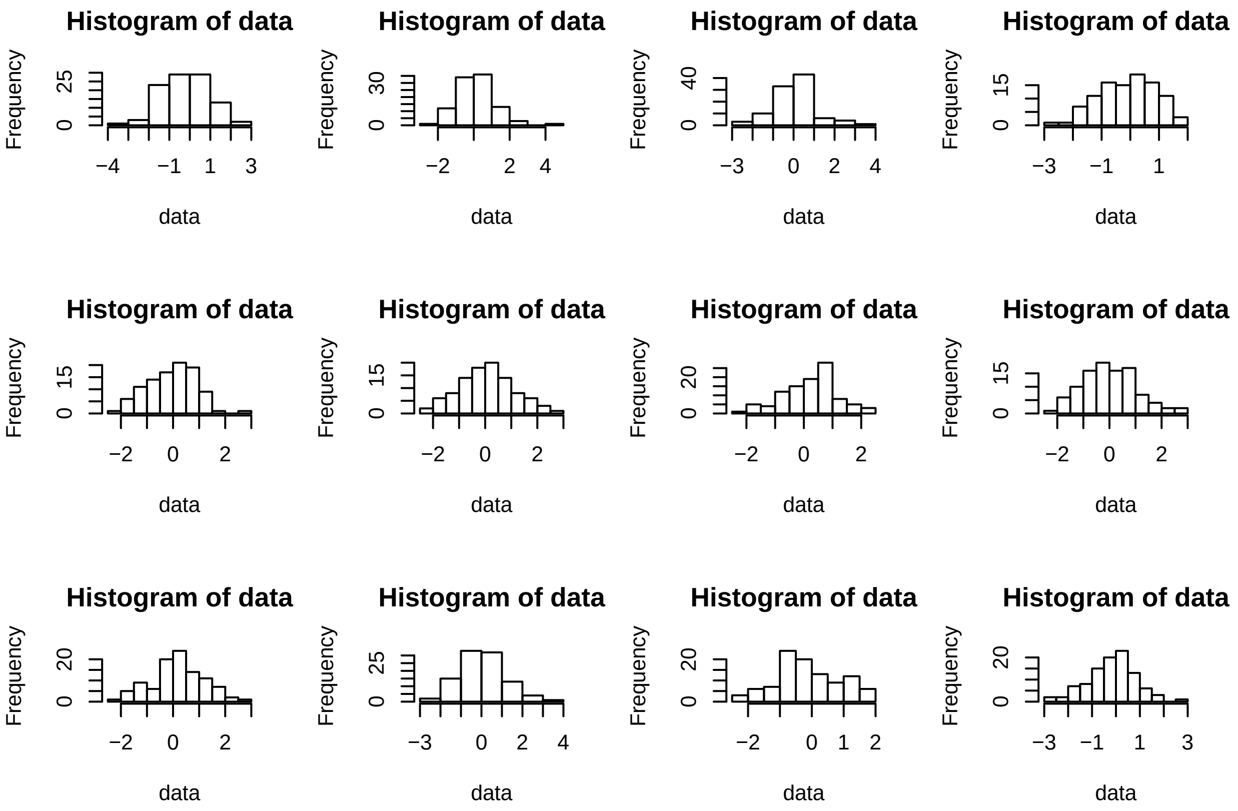


Do the samples appear to be normally distributed? Explain briefly. (Hand in the page with the twelve NPPs.)

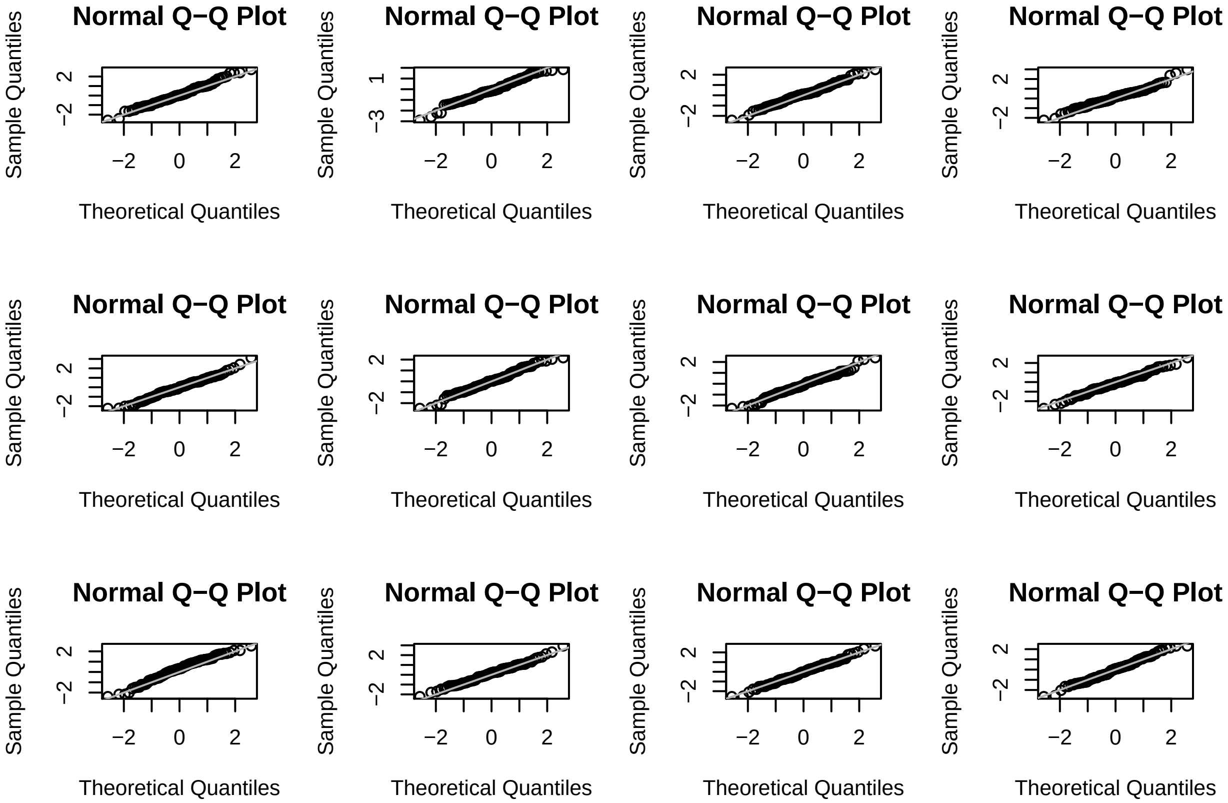
The samples seem normally distributed because they are pretty close to the line, but since the sample size is so small, it's hard to tell whether they really are actually normally distributed.

(c) Repeat parts (a) and (b) using twelve samples of 100 observations each. How does the appearance of the histograms and NPPs change with the increased sample size?

```r
#a
par(mfrow=c(3,4)) # plot 12 plots per page in a 4-by-3 array in row order
for (i in 1:12) {
  data <- rnorm(100, 0,1) # generate sample from Normal population with n = 10, mean = 0, SD = 1
  hist(data) # make histogram
}
```



```
#b
par(mfrow=c(3,4)) # plot 12 plots per page in a 4-by-3 array in row order
for (i in 1:12) {
  data <- rnorm(100, 0,1) # generate sample from Normal population with n = 10, mean = 0, SD = 1
  qqnorm(data) # make NPP
abline(0,1, col=gray(.7)) # add gray line with intercept = 0, slope = 1
}
```



The histograms, on average, look more normal and the NPPs look more normal as the sample size increases. The histograms achieve more of a bell shape and the NPP plot points get closer to the line.

(d) You can think of an entire NPP as being a statistic, since it is calculated from a sample. As such, it must have a sampling distribution, like any other statistic. Make a rough sketch (by hand) of the range in which you think 95% of NPPs would fall if sampled under these conditions, for both n = 10 and n = 100. Hand in these two pictures. (Think carefully about the shape of these ranges.)

5. Randomization Test (10 points)

Here we analyze data from the same study as in question 3. Women afflicted with PCOS often experience weight gain. Here we want to see if the PCOS subjects in our dataset have, on average, higher body mass index (BMI, a measure of weight that accounts for height) than control subjects. We will compare the two groups using a randomization test, as opposed to the traditional parametric analysis you did in question 3.

(a) Download the dataset bmi.csv from Moodle. Then read the dataset into R using the following command:

```
data <- read.csv("~/Desktop/school stuff/year2/spring20/stat21/bmi.csv", header=T)
```

(You may have to change the path to the file, depending on where you saved the downloaded file.) Verify that your dataset has 243 subjects with BMI in the first column and PCOS status in the second column (1 = PCOS, 99 = control). Define variables for PCOS and status as follows:

```
bmi <- data[,"bmi"]
status <- data[,"status"]
```

(b) Calculate the difference in the mean BMI for PCOS subjects and the control subjects using the following command:

```
mean_diff <- mean(bmi[status==1]) - mean(bmi[status==99]) # difference of group means
mean_diff
```

```
## [1] 3.431874
```

Difference between means: 3.432

(c) Now we will conduct a randomization test of the null hypotheses that PCOS subjects have the same average BMI as control subjects, against the one-sided alternative hypothesis that the PCOS subjects are higher. We will randomly permute the status labels 1000 times and observe the distribution of the test statistic (namely, the differences in the means of the two groups). First, define a vector diffs to hold the 1000 differences:

```
numshuffles <- 1000 # number of shuffles to run
diffs <- rep(NA, numshuffles) # create empty vector to hold shuffled differences
```

Now use a for loop to set up the 1000 randomizations:

Inside the for loop, you will need to randomly shuffle the status labels, calculate the difference in means using the shuffled group labels, and save the difference. To do the random shuffling, use the sample command:

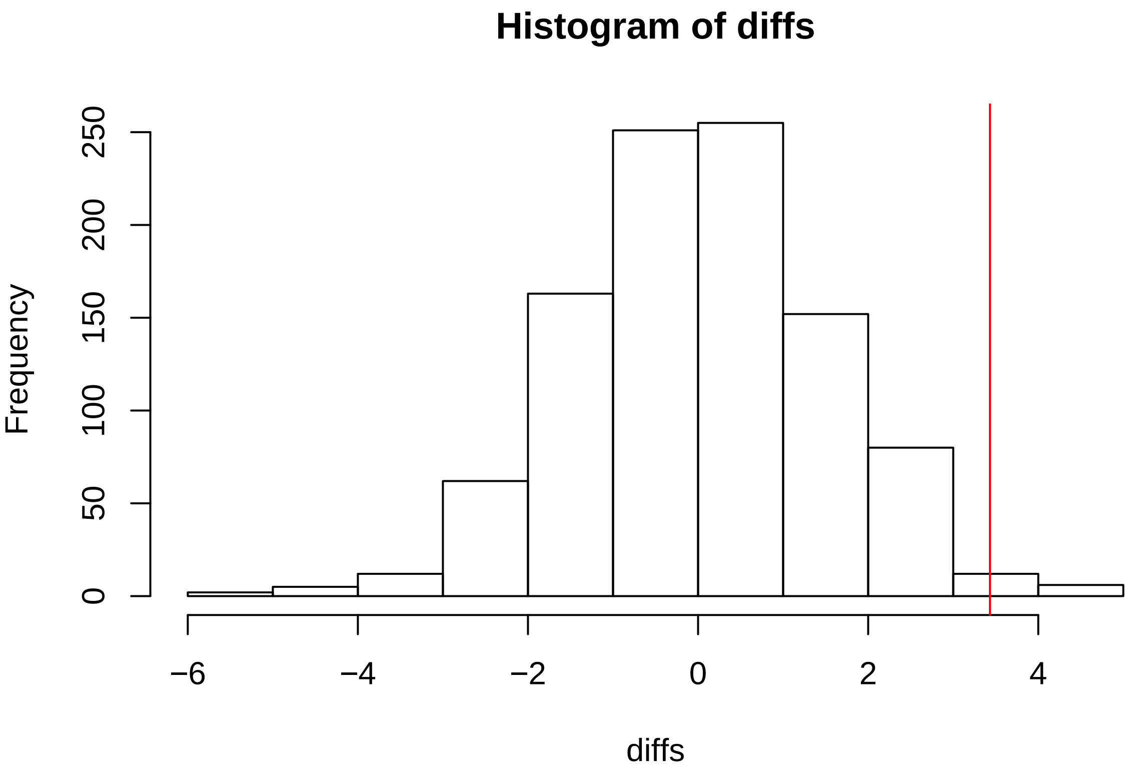
```
numshuffles <- 1000 # number of shuffles to run
diffs <- rep(NA, numshuffles) # create empty vector to hold shuffled differences

for(i in 1:numshuffles) {
  newstatus <- sample(status)
  diffs[i] <- mean(bmi[newstatus == 1]) - mean(bmi[newstatus == 99])
}
```

The other commands are left for you to figure out.

(d) Finally, once the loop has executed, make a histogram of the 1000 differences. Add a red vertical line showing the magnitude of the difference in the real dataset; you can do this using the abline command. Calculate what proportion of the differences from the shuffled datasets exceed the difference seen in the real dataset; this is your empirical p-value. What do you conclude about the null hypothesis?

```
hist(diffs)
abline(v = mean_diff, col = "red")
```



```
count <- 0
for (i in 1:numshuffles) {
  if (diffs[i] > mean_diff) {
    count <- count + 1
  }
}
count
```

```
## [1] 10
```

```
count/numshuffles #our empirical p-value
```

```
## [1] 0.01
```

Our p-value is less than 0.05, so we reject H_0. So, there is a significant difference between the "normal" systolic blood pressure and the blood pressure of the control subjects.

WHAT TO HAND IN: Hand in all of your R code, and your histogram showing the 1000 shuffled differences with the size of the real difference shown. (You can save R graphics windows as pdfs and then copy and paste them into a word processor document.) Also include your empirical p-value and your conclusion.

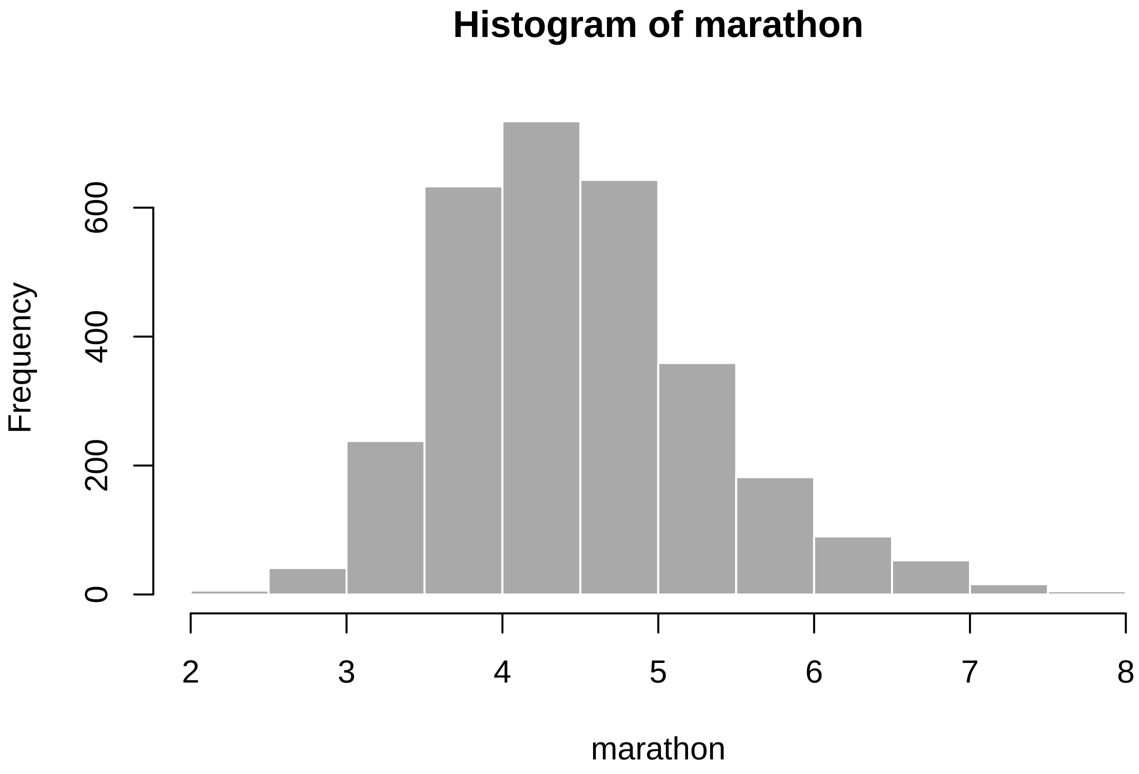
6. New York City Marathon (5 points)

The dataset nycmarathon.csv has finishing times of 3000 New York City marathon runners. Download the dataset from Moodle and read it into R using the following command:

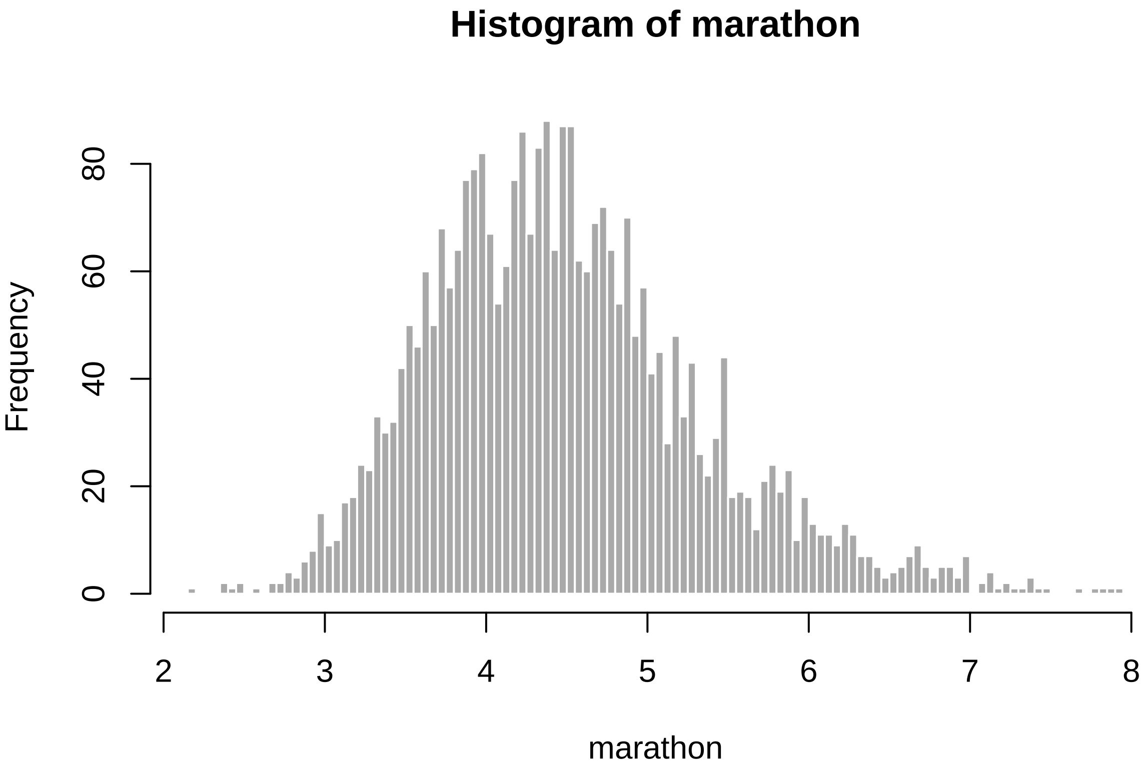
```r
marathon <- scan("~/Desktop/school stuff/year2/spring20/stat21/nycmarathon.csv")
```

Make a histogram using the following command:

```r
hist(marathon, col="darkgray", border="white", nclass=20)
```



```r
hist(marathon, col="darkgray", border="white", nclass=100)
```



It may be useful to change the nclass option to get more or fewer bars. Describe any interesting features of the dataset you discover.

The data is skewed right and I noticed that there are significant drops in people finishing near the hour marks, especially at hours 4 and 5, which might tell us that many of the runners ideally aim to finish before the exact hour mark.