lab4

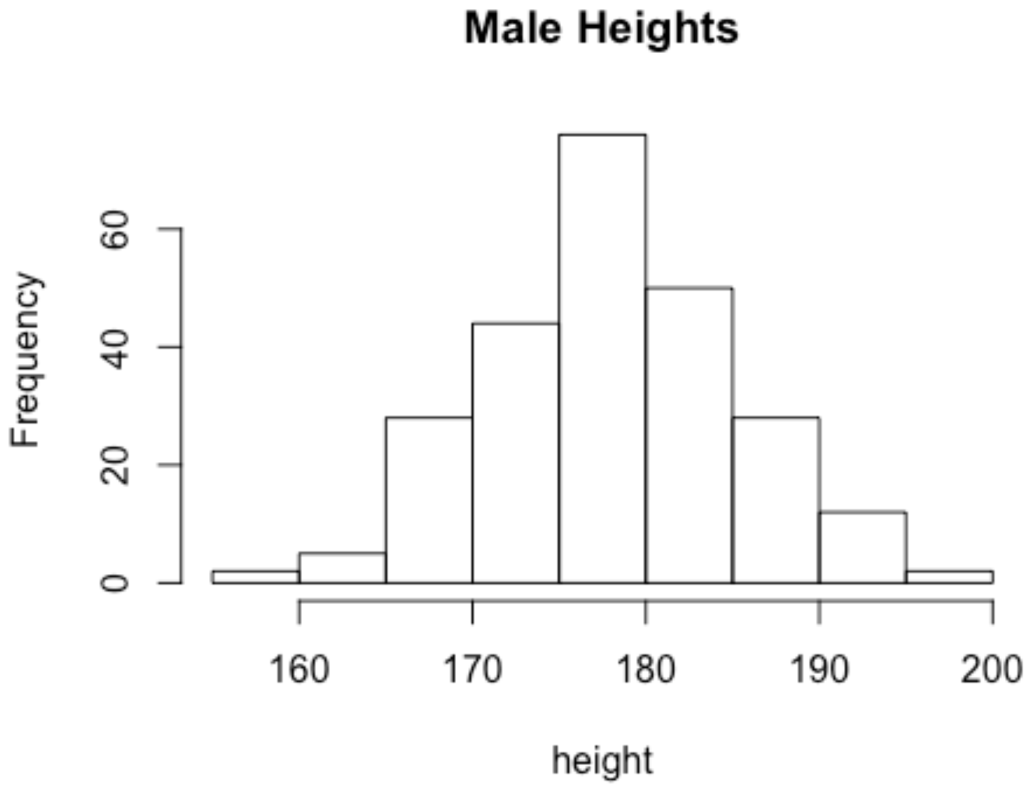
Christopher Andrews

11/4/2018

```
download.file("http://www.openintro.org/stat/data/bdims.RData", destfile =
"bdims.RData")
load("bdims.RData")

head(bdims)

##   bia.di bii.di bit.di che.de che.di elb.di wri.di kne.di ank.di sho.gi
## 1   42.9   26.0   31.5   17.7   28.0   13.1   10.4   18.8   14.1  106.2
## 2   43.7   28.5   33.5   16.9   30.8   14.0   11.8   20.6   15.1  110.5
## 3   40.1   28.2   33.3   20.9   31.7   13.9   10.9   19.7   14.1  115.1
## 4   44.3   29.9   34.0   18.4   28.2   13.9   11.2   20.9   15.0  104.5
## 5   42.5   29.9   34.0   21.5   29.4   15.2   11.6   20.7   14.9  107.5
## 6   43.3   27.0   31.5   19.6   31.3   14.0   11.5   18.8   13.9  119.8
##   che.gi wai.gi nav.gi hip.gi thi.gi bic.gi for.gi kne.gi cal.gi ank.gi
## 1   89.5   71.5   74.5   93.5   51.5   32.5   26.0   34.5   36.5   23.5
## 2   97.0   79.0   86.5   94.8   51.5   34.4   28.0   36.5   37.5   24.5
## 3   97.5   83.2   82.9   95.0   57.3   33.4   28.8   37.0   37.3   21.9
## 4   97.0   77.8   78.8   94.0   53.0   31.0   26.2   37.0   34.8   23.0
## 5   97.5   80.0   82.5   98.5   55.4   32.0   28.4   37.7   38.6   24.4
## 6   99.9   82.5   80.1   95.3   57.5   33.0   28.0   36.6   36.1   23.5
##   wri.gi age  wgt   hgt sex
## 1   16.5  21 65.6 174.0   1
## 2   17.0  23 71.8 175.3   1
## 3   16.9  28 80.7 193.5   1
## 4   16.6  23 72.6 186.5   1
## 5   18.0  22 78.8 187.2   1
## 6   16.9  21 74.8 181.5   1

mdims <- subset(bdims, bdims$sex == 1)
fdims <- subset(bdims, bdims$sex == 0)
```

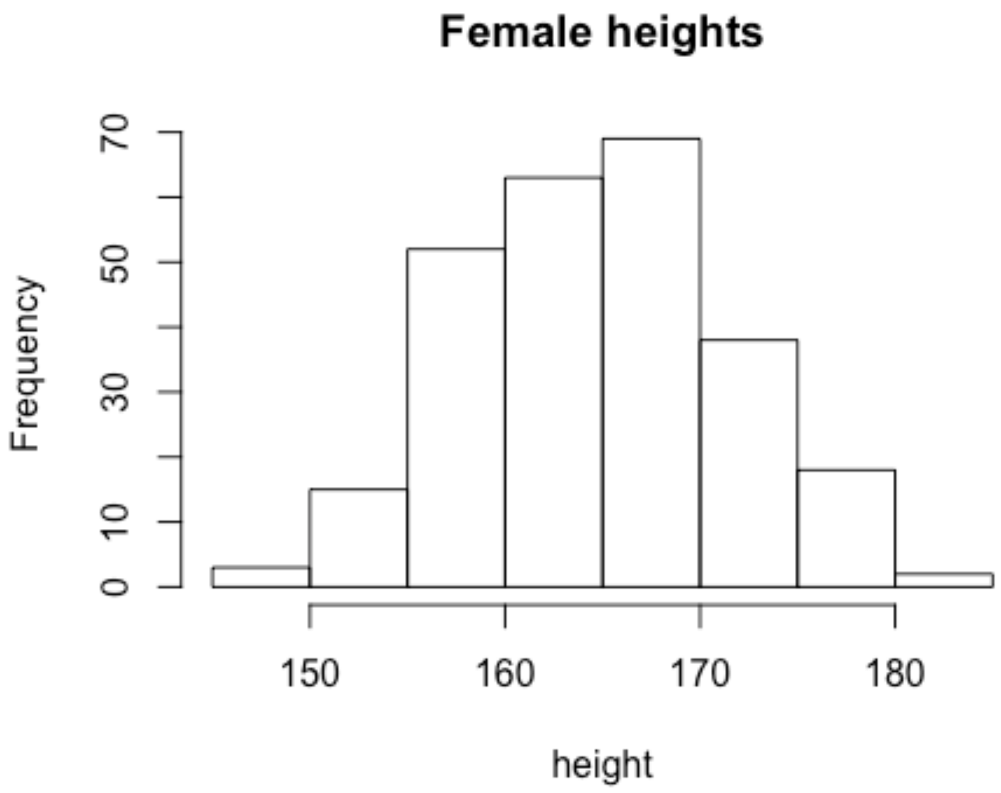
➡️ Exercise 1: Generate separate histograms of the men's and women's heights. Then, compare and contrast the center, spread, and shape of these two height distributions. (Hint: It would be advisable to also generate summary statistics so that you can quantify the center and spread of these distributions.) For the males height distribution, the spread is symmetric and without skew. The center lies between 175cm and 180cm. This is displayed in the summary statistics that show that the median lies at 177.8 cm. Its symmetry is also displayed by q1 and q3 calculations being an equal amount of points away from the median, q1 at 4.8 away and q3 at 4.8 away from the median as well. For the female distribution is slightly skewed to the left, but could be seen as symmetric. the center resides centimeters, which sounds right because on average females are shorter than males. Its ever so slight skewness is showed by Median-1q = 4.5 and 3q-median = 5 calculations. the median lies slightly closer to the first quartile.

```
hist(mdims$hgt, main = "Male Heights", xlab = "height" )
```

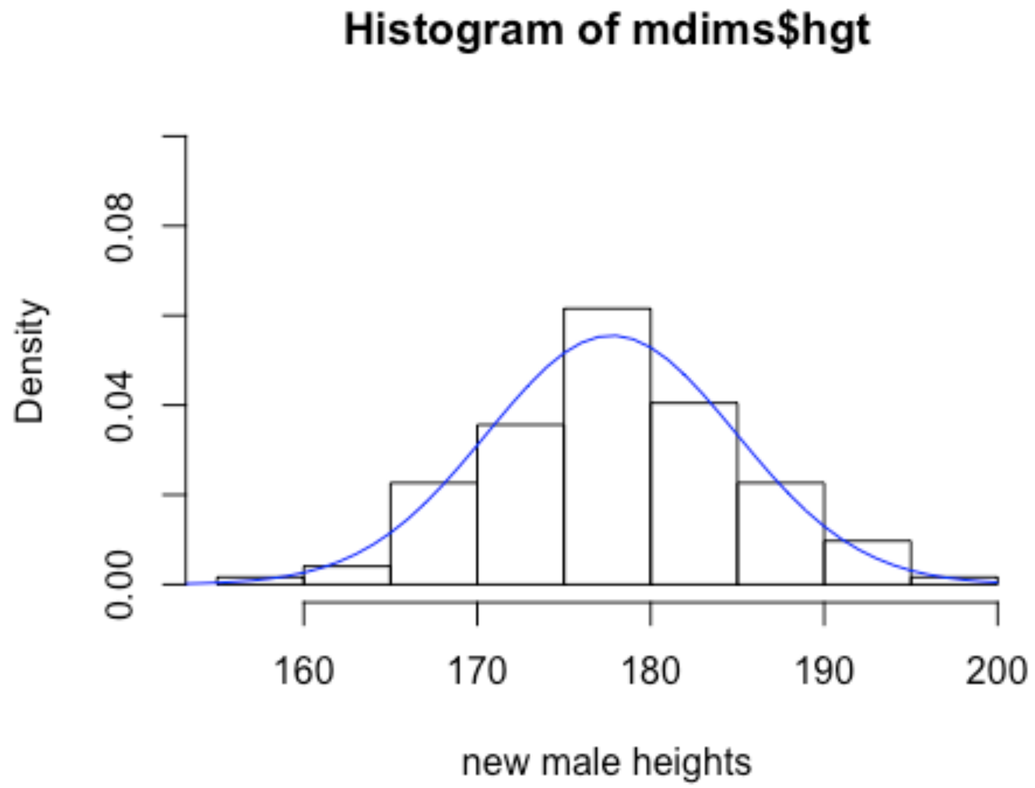


```
summary(mdims$hgt)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   157.2   172.9   177.8   177.7   182.7   198.1
```

```
hist(fdims$hgt, main = "Female heights", xlab = "height")
```

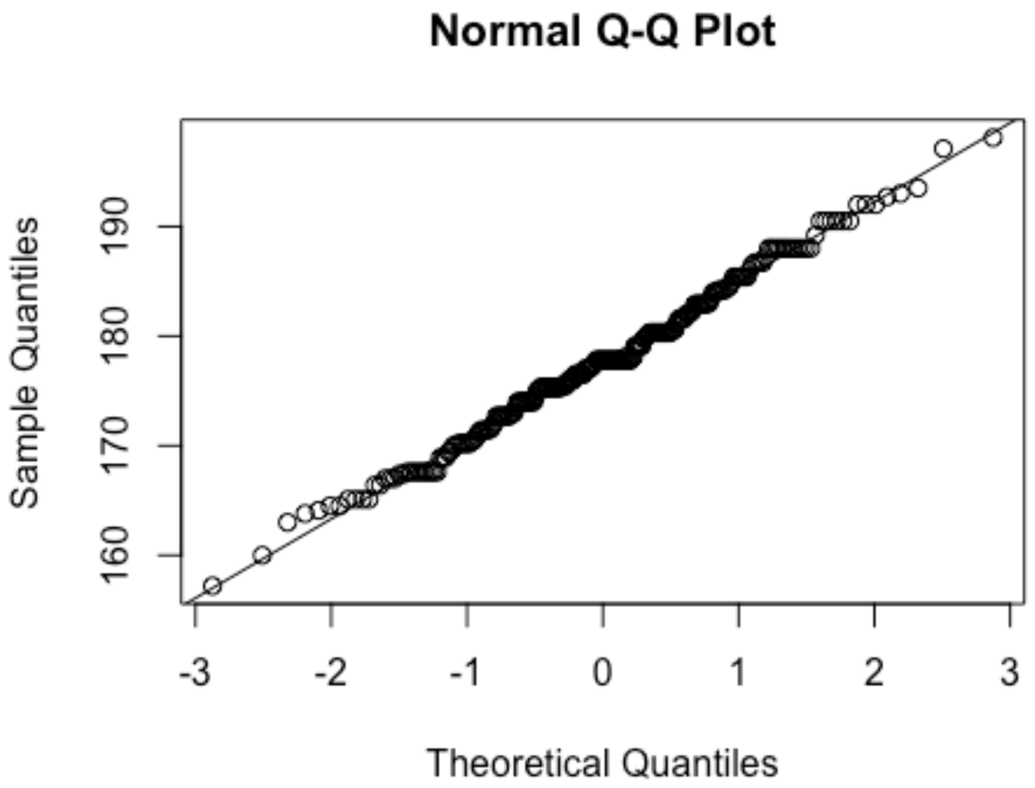


```
summary(fdims$hgt)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   147.2   160.0   164.5   164.9   169.5   182.9

mhgtmean <- mean(mdims$hgt)
mhgtsd <- sd(mdims$hgt)

hist(mdims$hgt, probability = TRUE, ylim = c(0,0.1), xlab = "new male heights")
x <- 150:200
y <- dnorm(x = x, mean = mhgtmean, sd = mhgtsd)
lines(x = x, y = y, col = "blue")
```


Histogram of mdims$hgt
Density
0.00
0.04
0.08
160
170
180
190
200
new male heights

Exercise 2: Based on this plot, does it appear that the men's height data follow a nearly normal distribution? Explain. Based on this visual there is evidence to suggest that it follows a close to normal distribution.

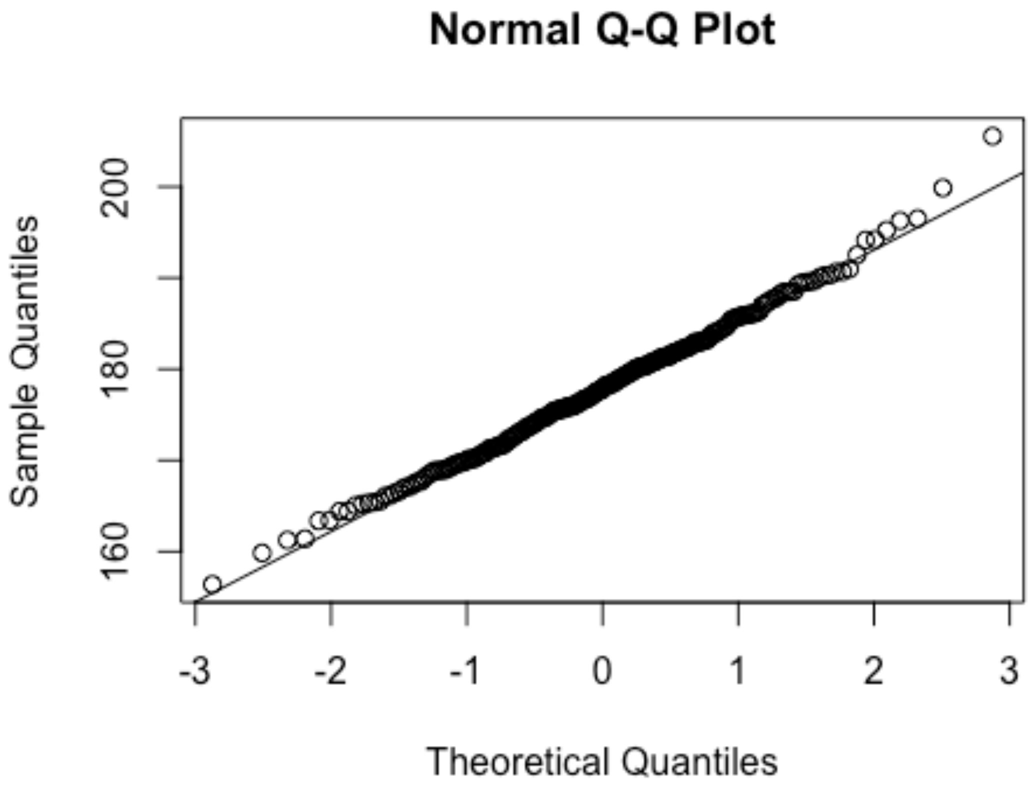
```
qqnorm(mdims$hgt)
qqline(mdims$hgt)
```



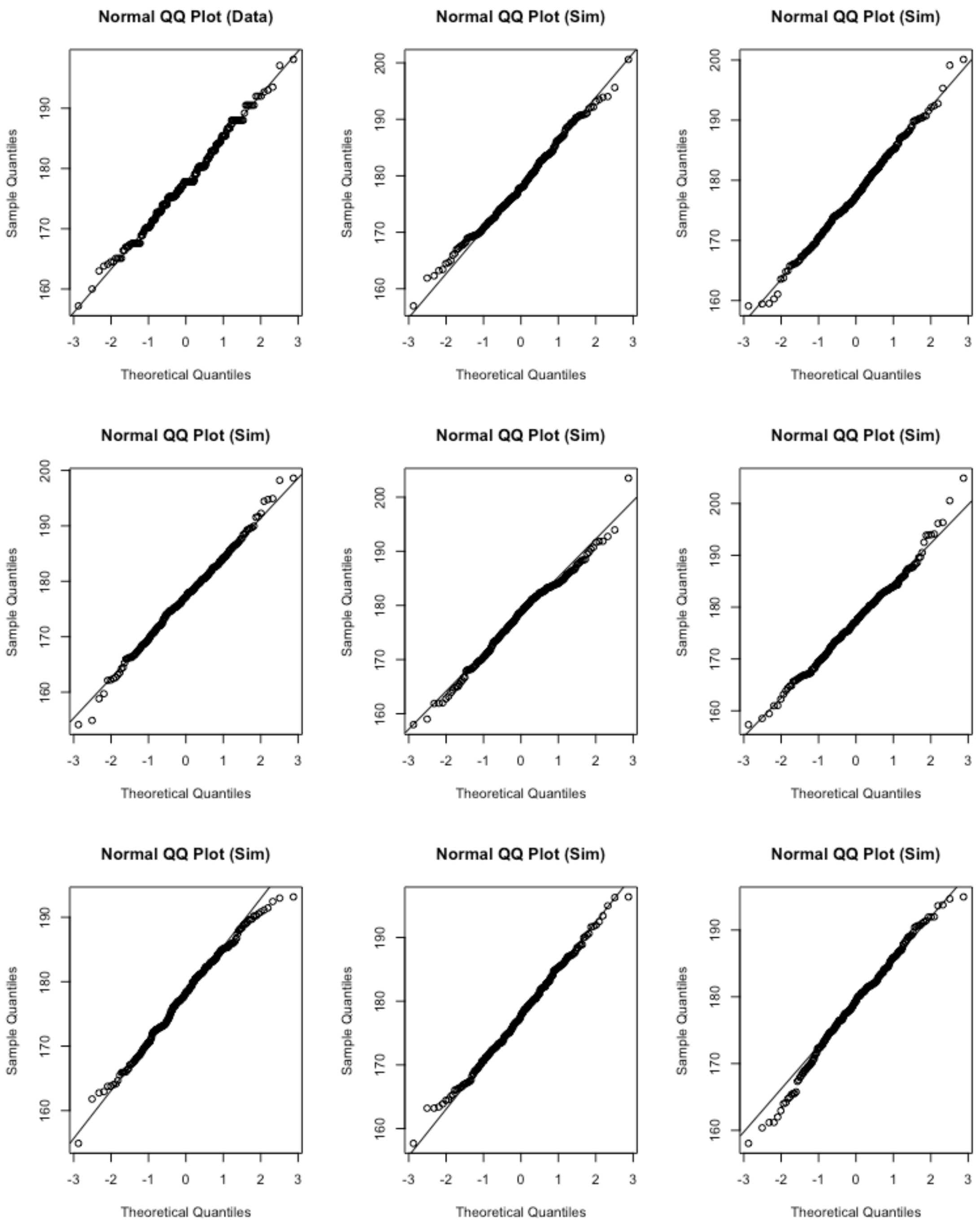
```
sim_norm <- rnorm(n = length(mdims$hgt), mean = mhgtmean, sd = mhgtsd)
```

Exercise 3: Make a normal (Q-Q) probability plot of sim_norm. Do all of the points fall on the line? How does this plot compare to the normal probability plot for the actual data? No, not all the points fall on the line, there are some deviations. But the trend follows the line pretty closely, for the most part and it is safe to assume that this simulation (sim_norm) represents a normal distribution. Not so suprisingly, this line is somewhat close to the actual data, seen above.

```r
qqnorm(sim_norm)
qqline(sim_norm)
```



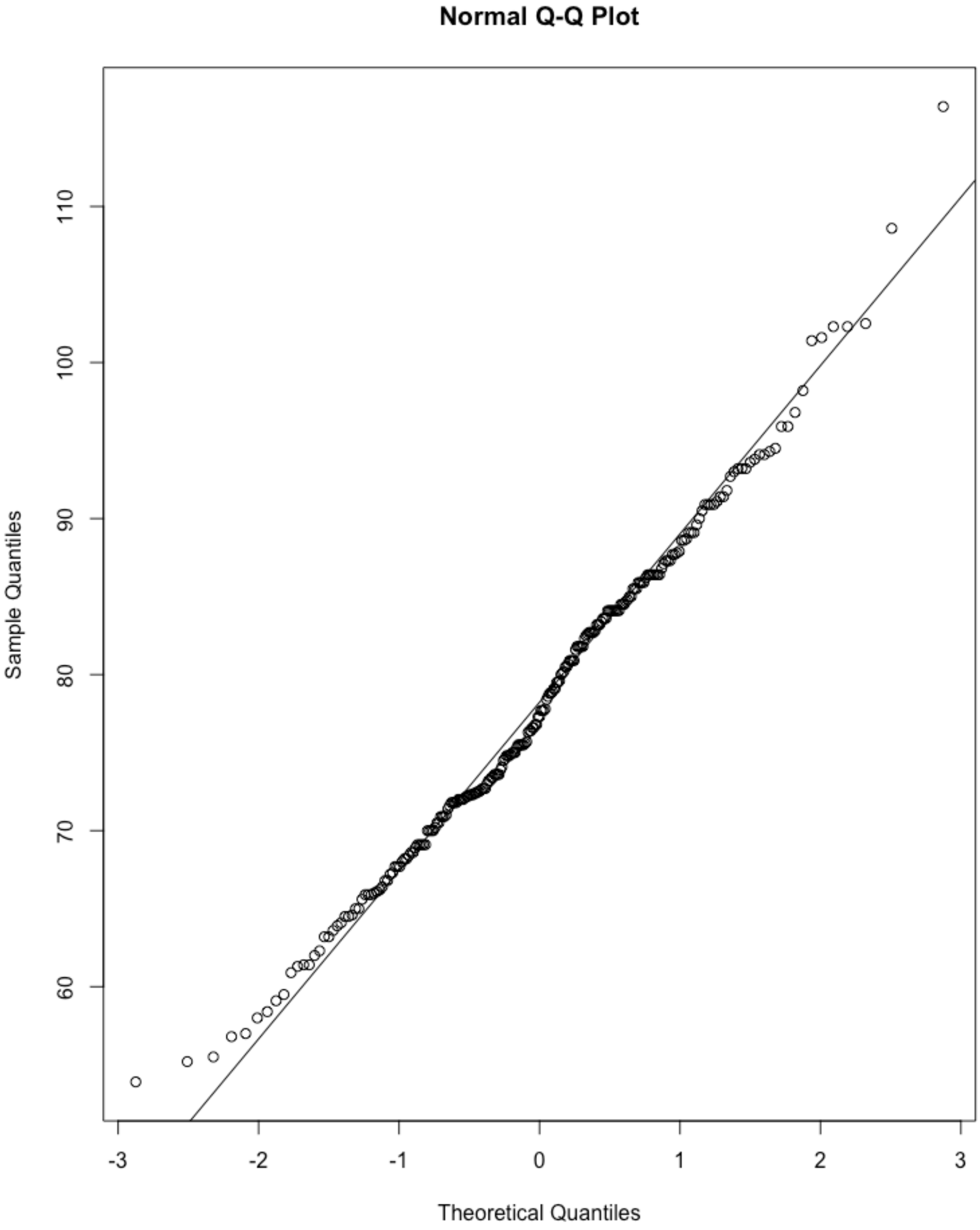
```
qqnormsim(mdims$hgt)
```



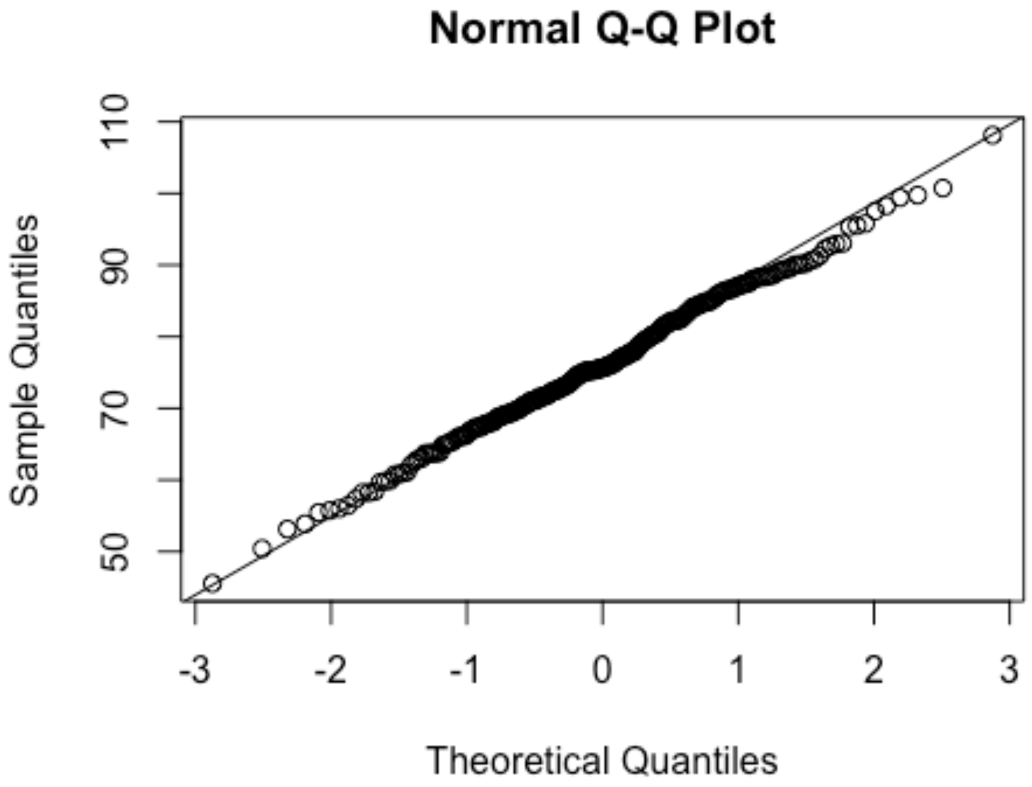
Exercise 4: Does the normal probability plot for mdims$hgt look similar to the plots created for the simulated data? That is, do the plots provide evidence that the male heights are nearly normal? Explain. Yes, the QQplot with the actual data does look very similar to the simulated ones. Some look more simulated than others, but that is almost guaranteed when doing simulations, some will be a more accurate representation of the actual data than others.

Exercise 5: Using the same procedure you used to judge the normality of the male height data in Exercises 2 through 4, explain your judgment as to whether or not the male weights appear to come from a normal distribution. They do appear to come from a normal distribution, but in the simulation, the plot is more densly populated on the theoretical line as it gets closer to 0 on the x-axis.

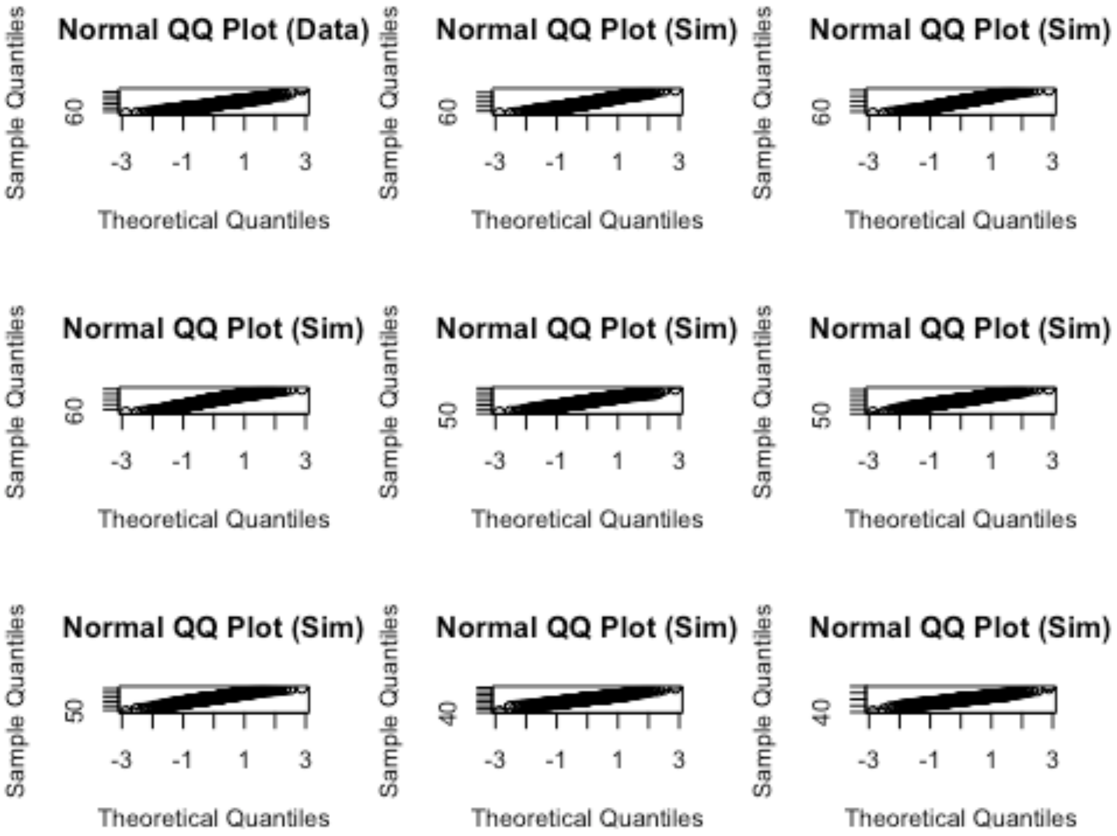
```r
qqnorm(mdims$wgt)
qqline(mdims$wgt)
```


Normal Q-Q Plot
Sample Quantiles
60
70
80
90
100
110
Theoretical Quantiles
-3
-2
-1
0
1
2
3

```
mwgtmean <- mean(mdims$wgt)
mwgtsd <- sd(mdims$wgt)
sim_norm_wgt <- rnorm(n = length(mdims$wgt), mean = mwgtmean, sd = mwgtsd)
qqnorm(sim_norm_wgt)
qqline(sim_norm_wgt)
```



```
qqnormsim(mdims$wgt)
```



```
1 - pnorm(q = 182, mean = mhgtmean, sd = mhgtsd)

## [1] 0.2768345

sum(mdims$hgt > 182)/length(mdims$hgt)

## [1] 0.2631579
```

Exercise 6: Write out two probability questions that you would like to answer - one regarding male heights and one regarding male weights. Calculate those probabilities using both the theoretical normal distribution as well as the empirical distribution (four probabilities in all). Which variable, height or weight, is closer to normal? Explain your reasoning by comparing each empirical distribution to the corresponding theoretical normal distribution Question 1. Male heights greater than 6 feet

```
SixFtInCM <- 72*2.54 #convert to CM
maleMean <- mean(mdims$hgt)
maleSD <- sd(mdims$hgt)

1-pnorm(q = SixFtInCM, mean = maleMean, sd = maleSD)

## [1] 0.237375

sum(mdims$hgt > SixFtInCM)/length(mdims$hgt)

## [1] 0.2510121
```

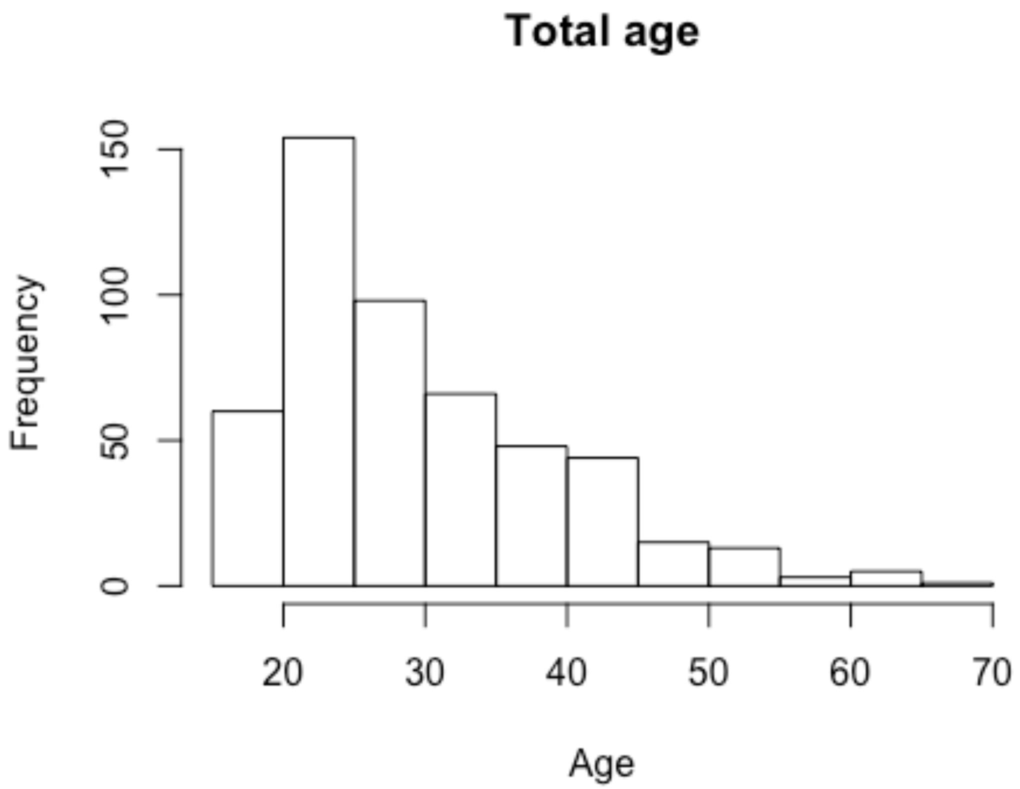
We are able to find a close approximation through both of these measures because males heights are distributed normally. It shows in this sample about 25 percent of men are taller than 6ft tall. This is normal, which is shown by the closeness of the two calculations.

Question 2. male weights less than 155 lbs.

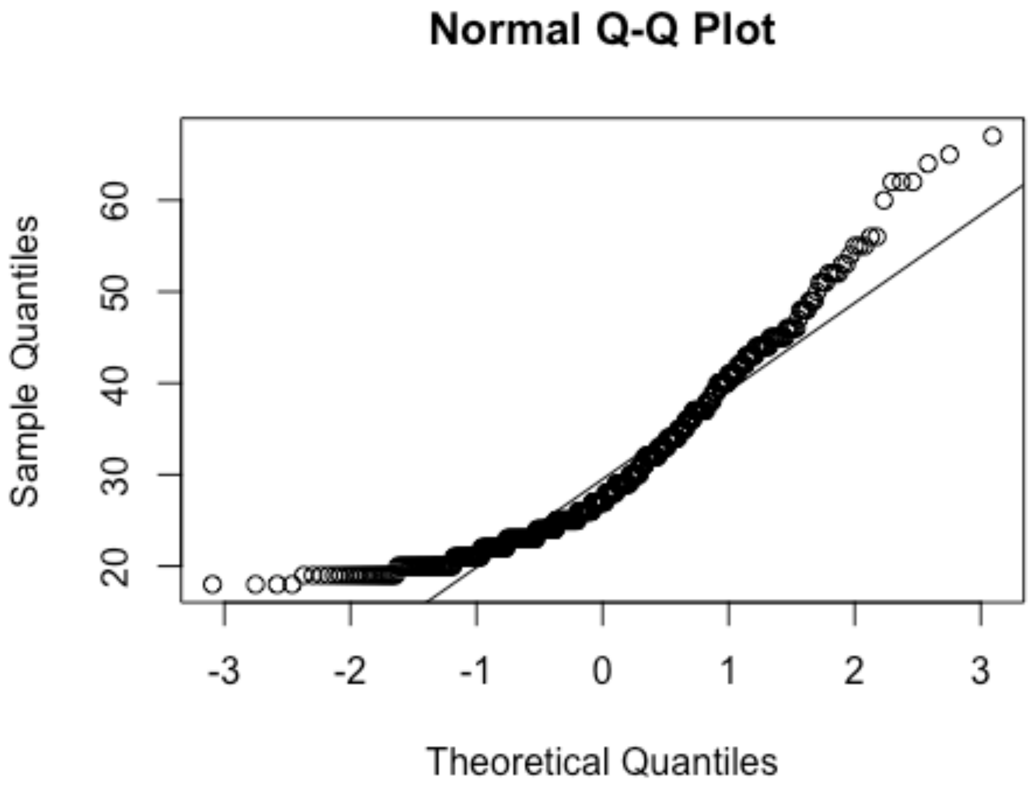
```
WeightInKG <- 155/2.205 #convert to kg
maleWgtMean <- mean(mdims$wgt)
maleWgtSD <- sd(mdims$wgt)

pnorm(q = WeightInKG, mean = maleWgtMean, sd = maleWgtSD)

## [1] 0.2276288

sum(mdims$wgt < WeightInKG)/length(mdims$wgt)

## [1] 0.2307692
```

Once again, we are able to find a close approximation through these measures with regards to male weights, in this case it was mens weights who are less than 155 lbs, and in this sample that comes to about 23 percent. This is normal, which is seen by the clseness of the two calculations. Homework Assignment 1. Now let's consider some of the other variables in the body dimensions data set. Using the figures on the next page, match each histogram to its normal probability plot. All of the variables have been standardized (by first subtractig the mean, and then dividing by the standard deviation), so the units won't be of any help. If you are uncertain based on these figures, you can generate the plots in R to check. 1) D 2) A 3) B 4) C

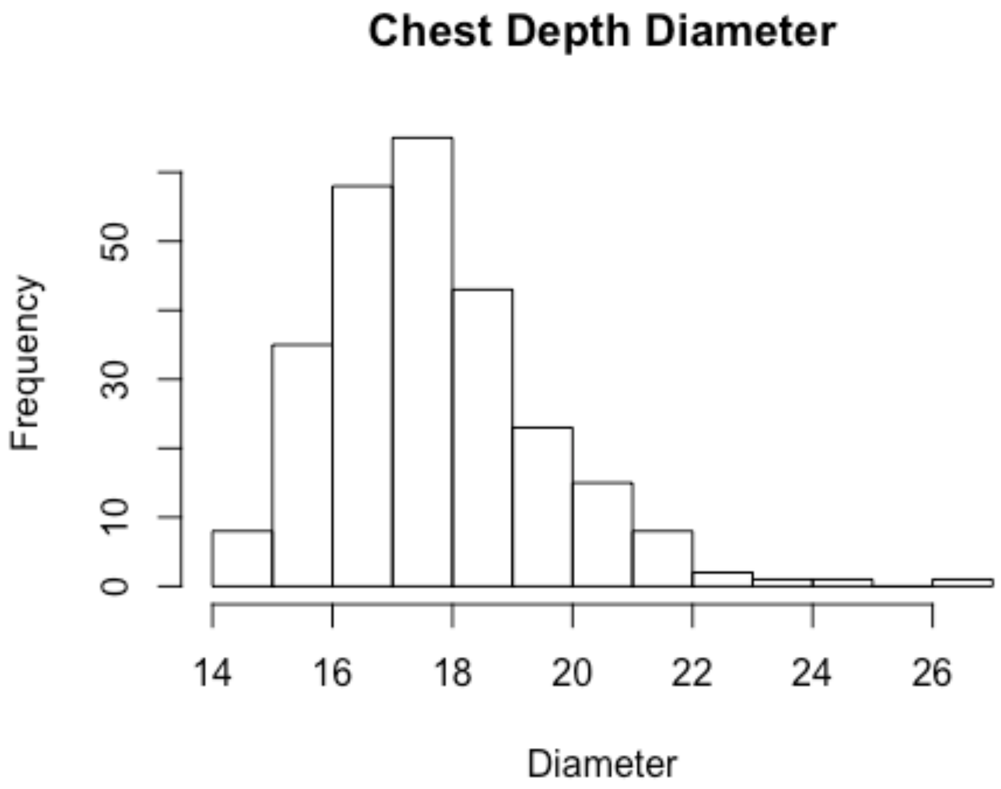
```r
hist(bdims$age, main = "Total age", xlab = "Age")
```



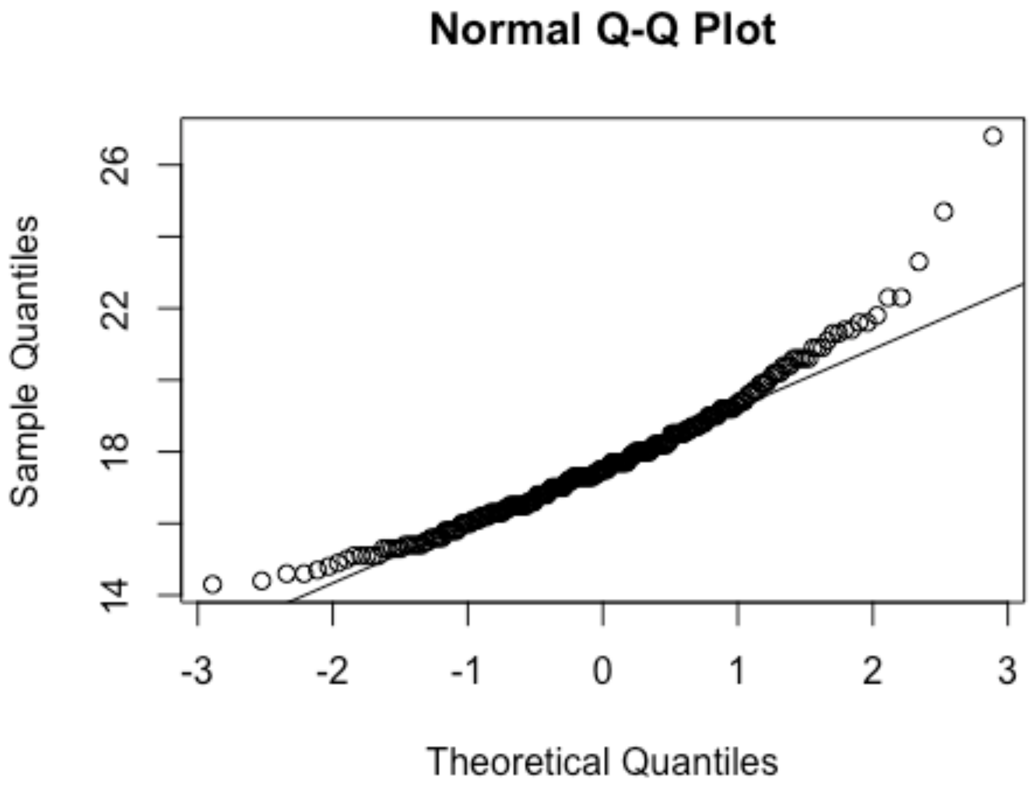
```r
qqnorm(bdims$age)
qqline(bdims$age)
```



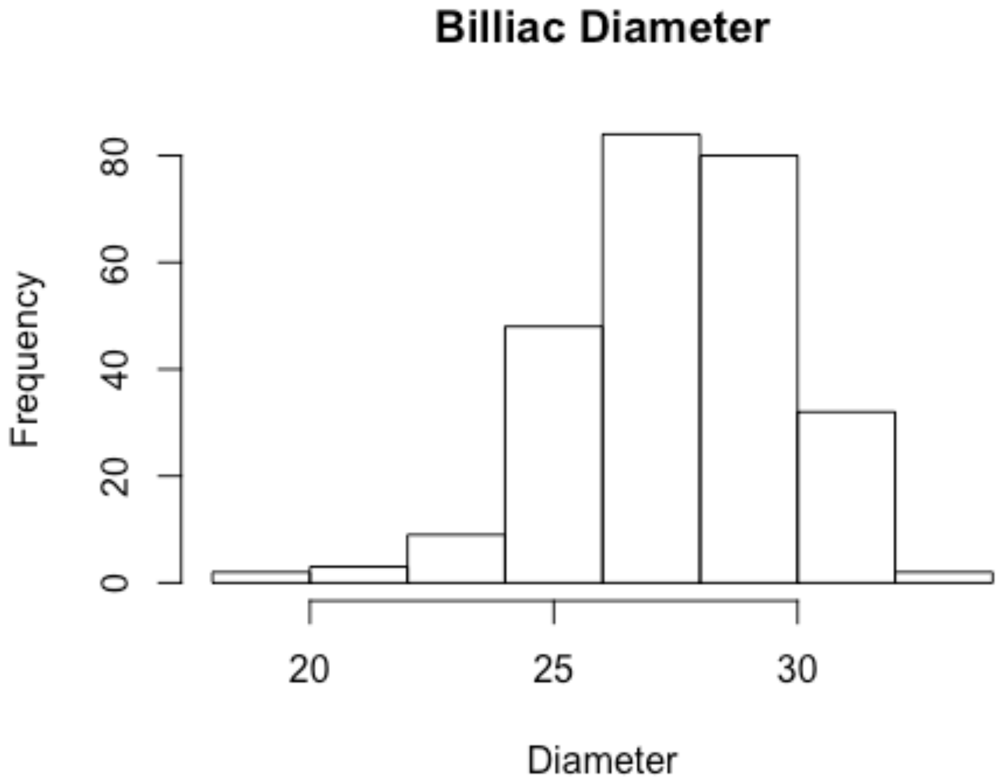
```r
hist(fdims$che.de, main = "Chest Depth Diameter", xlab = "Diameter")
```



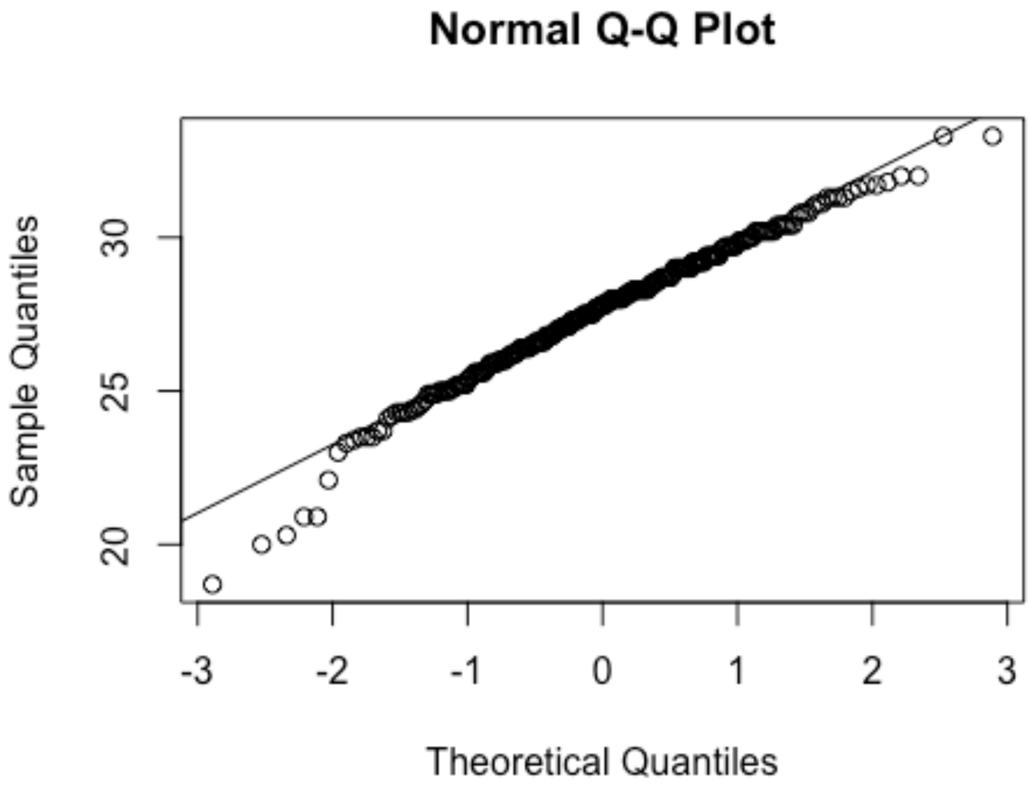
```r
qqnorm(fdims$che.de)
qqline(fdims$che.de)
```



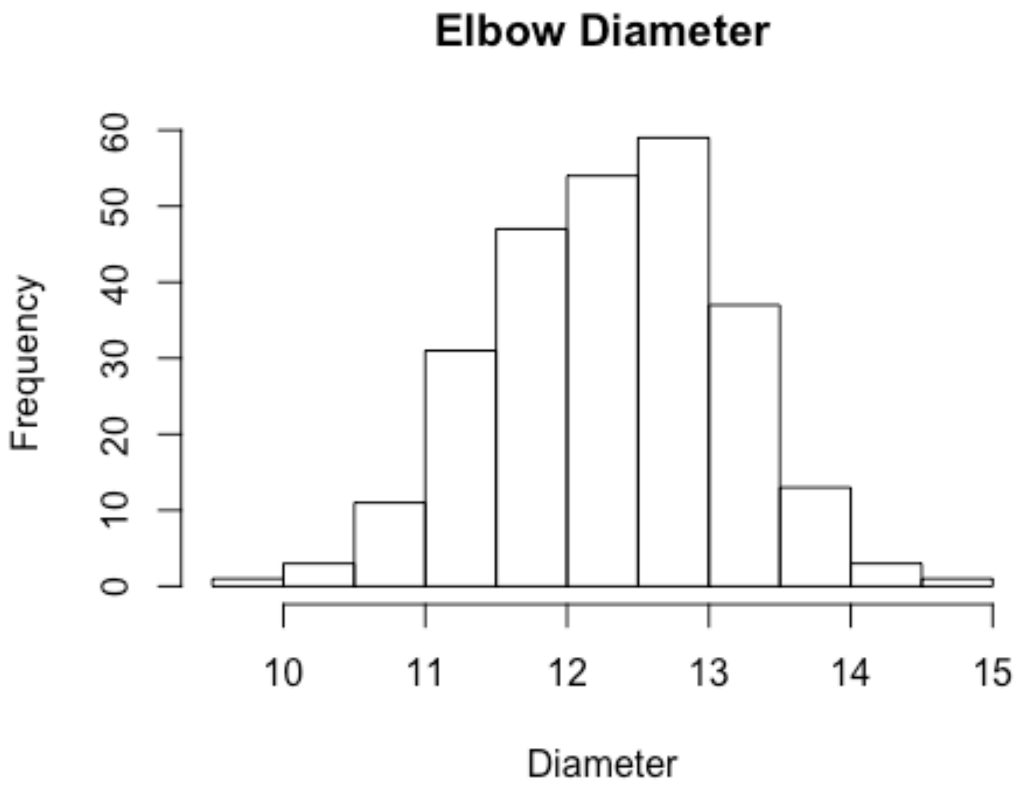
```
hist(fdims$bii.di, main = "Billiac Diameter", xlab = "Diameter")
```



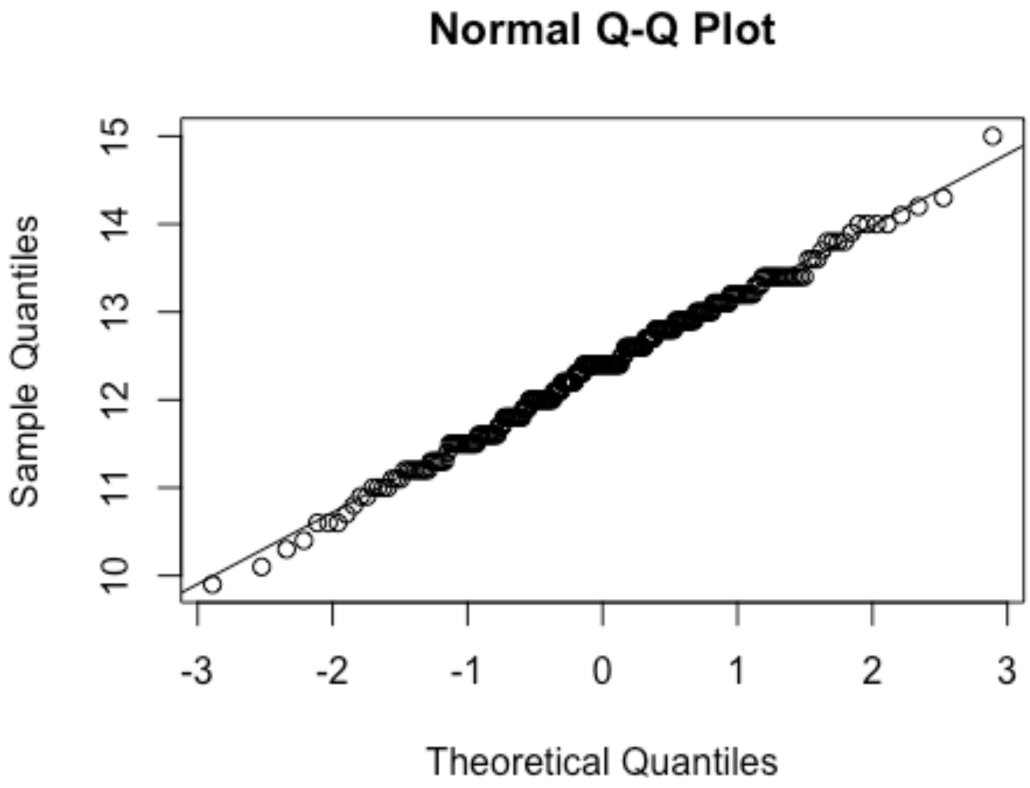
```r
qqnorm(fdims$bii.di)
qqline(fdims$bii.di)
```



```r
hist(fdims$elb.di, main = "Elbow Diameter", xlab = "Diameter")
```



```
qqnorm(fdims$elb.di)
qqline(fdims$elb.di)
```

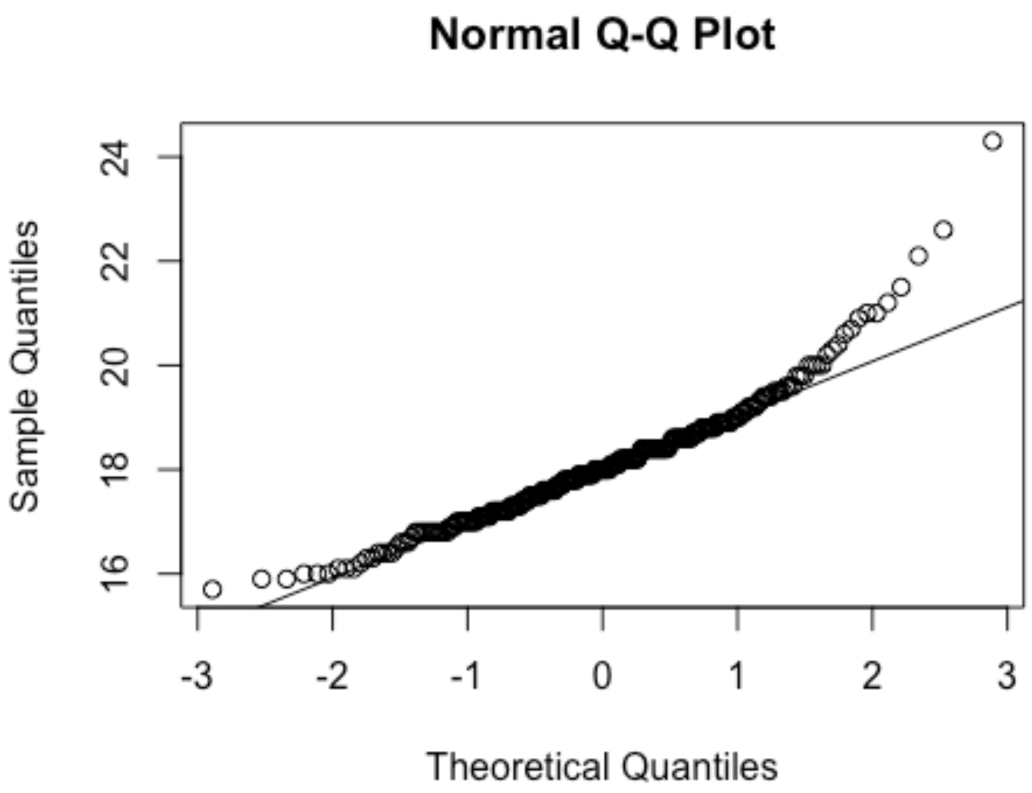


2. Note that normal probability plot D has a slight stepwise pattern. Why do you think this is the case?

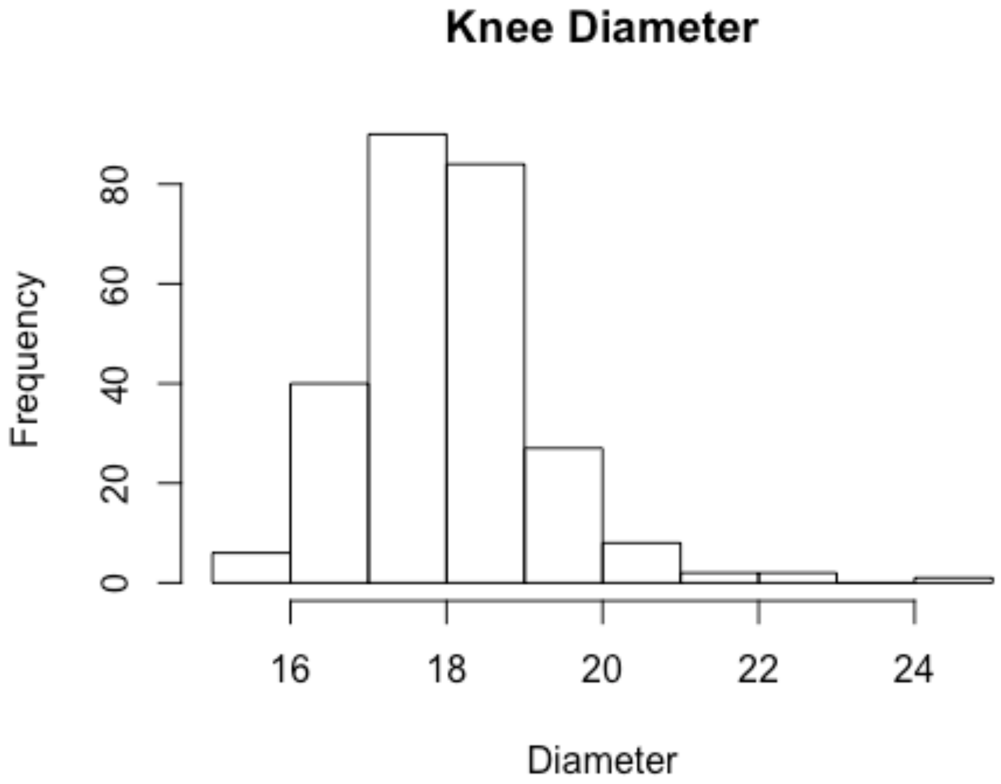
Because in this case the variable we are looking at is discrete in the way it is given to us. The variable (age) is a continuous variable because every second/millisecond/nanosecond etc we are getting older. But in this case the age is just give in years and is represented as a discrete variable. If we were getting it in continuous form then it would be less of a stair step pattern.

3. As you can see, normal probability plots can be used both to assess normality and visualize skewness. Produce a normal probability plot for female knee diameter (kne.di). Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Explain your reasoning. Use a histogram to confirm your findings. The plot of this variable is right skewed. You can see this by the plot on the histogram and the shape of it, but also through the normal QQ plot, which shows right skewness but how it tailers off the normal line on the right side of the 0 on the x-axis.

```r
qqnorm(fdims$kne.di)
qqline(fdims$kne.di)
```



```r
hist(fdims$kne.di, main = "Knee Diameter", xlab = "Diameter")
```


Knee Diameter
Frequency
Diameter