Lab3

Bin Lin

2016-9-23

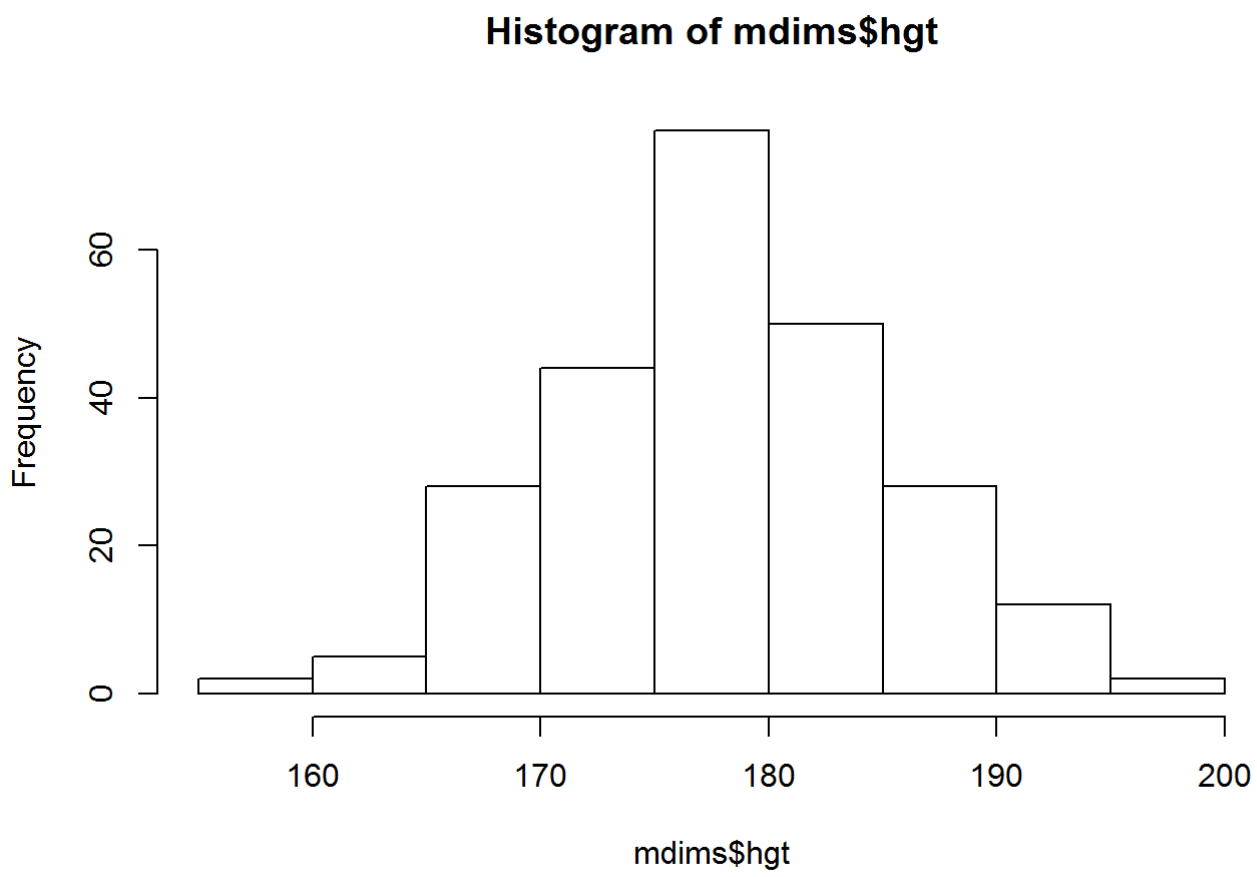
```
download.file("http://www.openintro.org/stat/data/bdims.RData", destfile = "bdims.RData")
load("bdims.RData")
head(bdims)
```

```
##   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
```

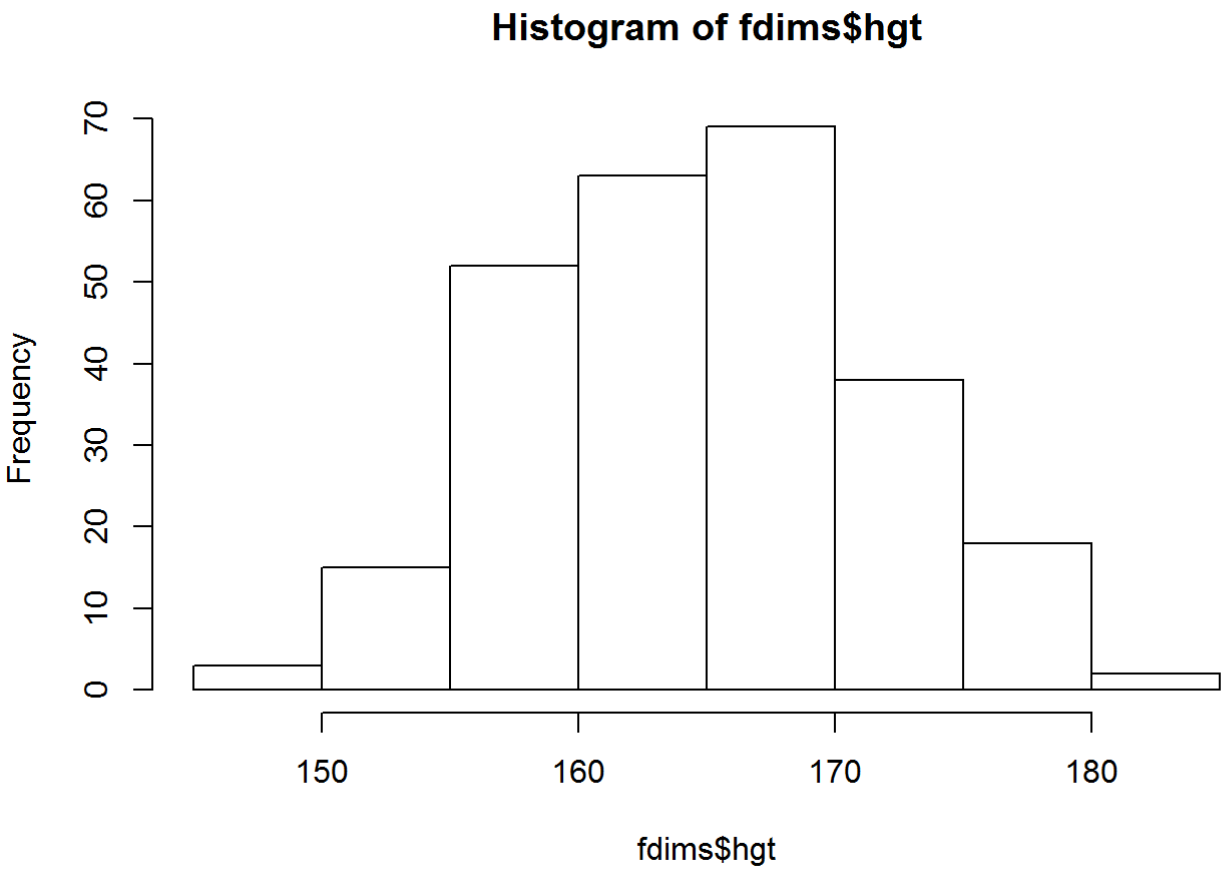
```
mdims <- subset(bdims, sex == 1)
fdims <- subset(bdims, sex == 0)
```

Excercise 1: Make a histogram of men's heights and a histogram of women's heights. How would you compare the various aspects of the two distributions?

```
hist(mdims$hgt)
```



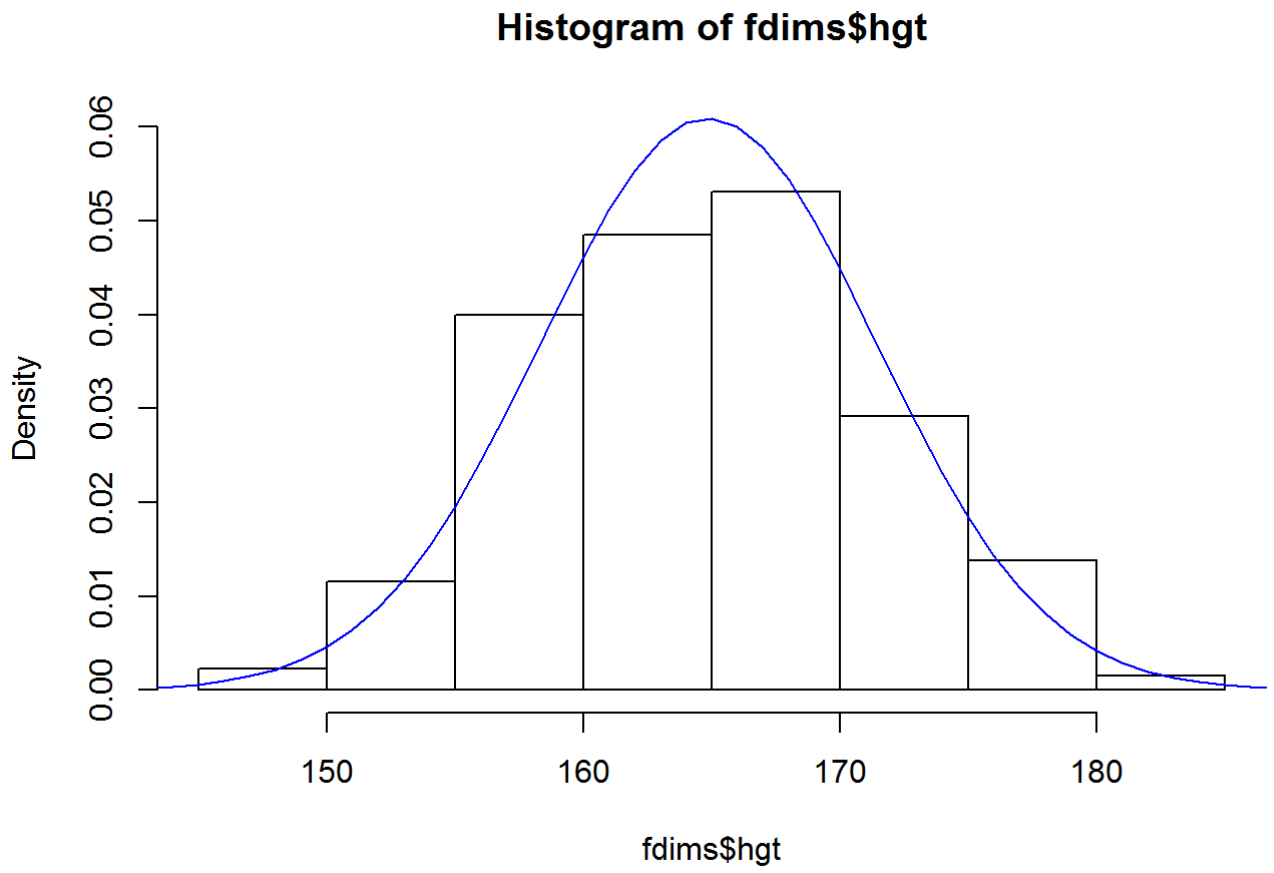

```
hist(fdims$hgt)
```



Both figures are bell-shaped with just one peak. But the peak for the male is more towards the center, while for female the peak is more widely spread.

Exercise 2: Based on the this plot, does it appear that the data follow a nearly normal distribution?

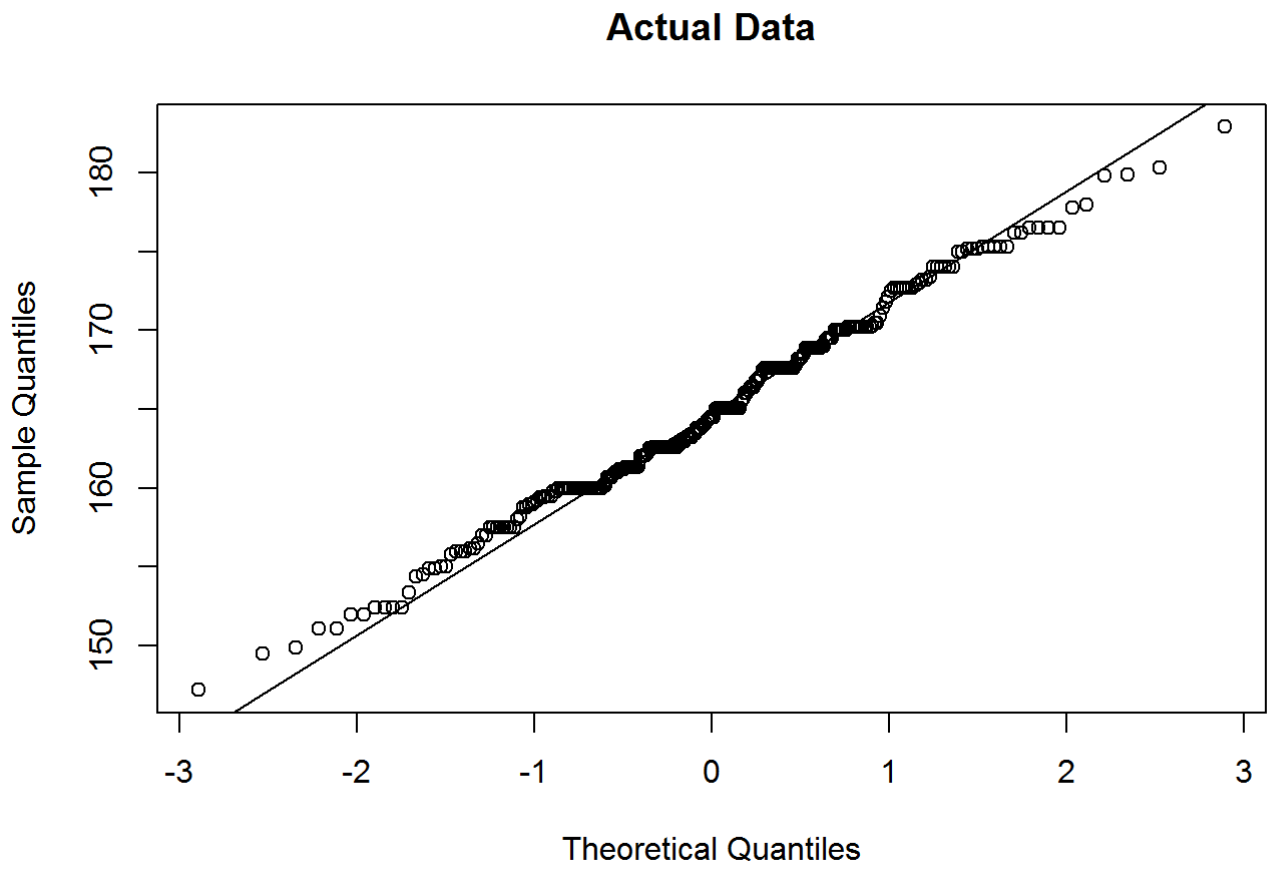
```
fhgtmean <- mean(fdims$hgt)
fhgtsd   <- sd(fdims$hgt)
hist(fdims$hgt, probability = TRUE, ylim = c(0, 0.06))
x <- 140:190
y <- dnorm(x = x, mean = fhgtmean, sd = fhgtsd)
lines(x = x, y = y, col = "blue")
```



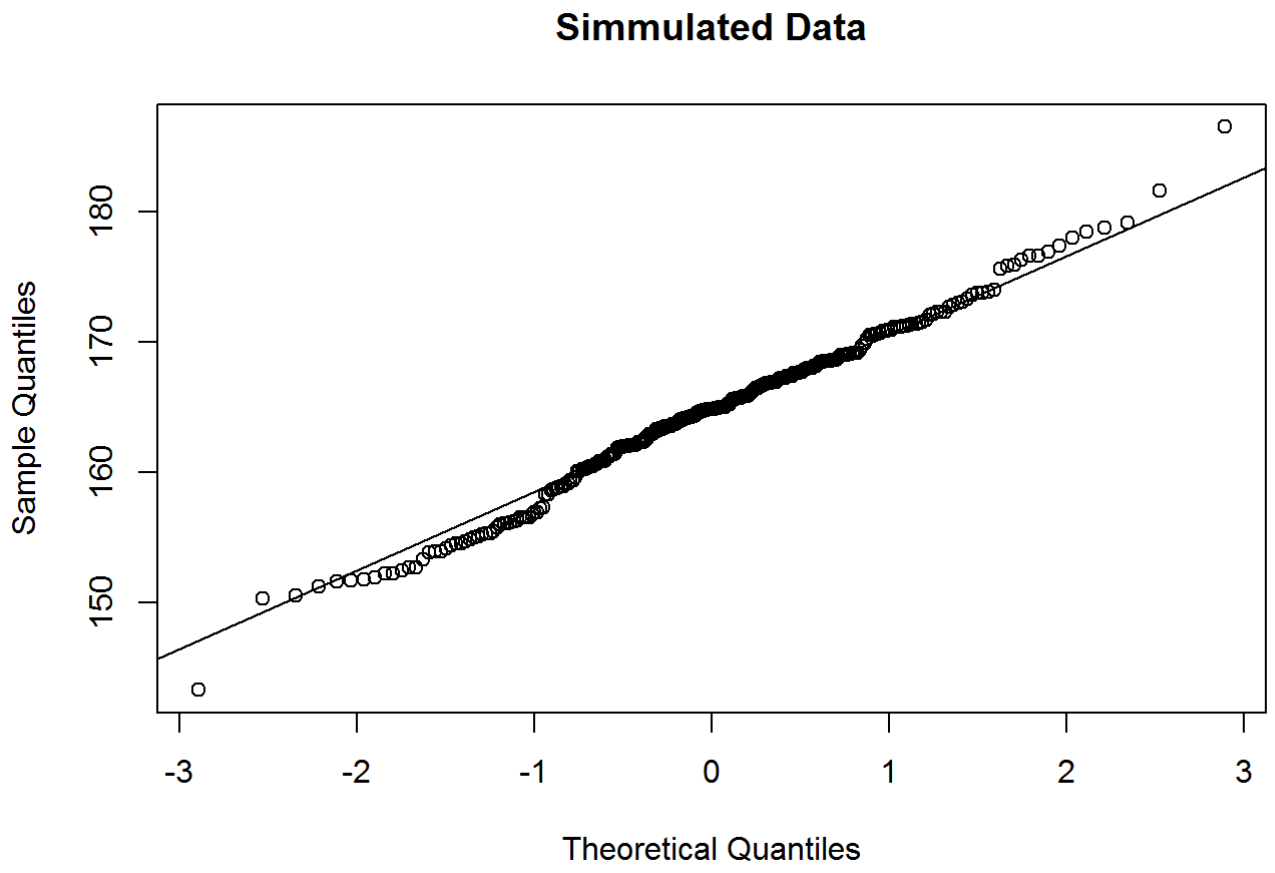
According to this graph, the distribution of women's height is closed to normal, but not quite as the peak appears to be flatter.

Exercise 3: Make a normal probability plot of sim_norm. Do all of the points fall on the line? How does this plot compare to the probability plot for the real data

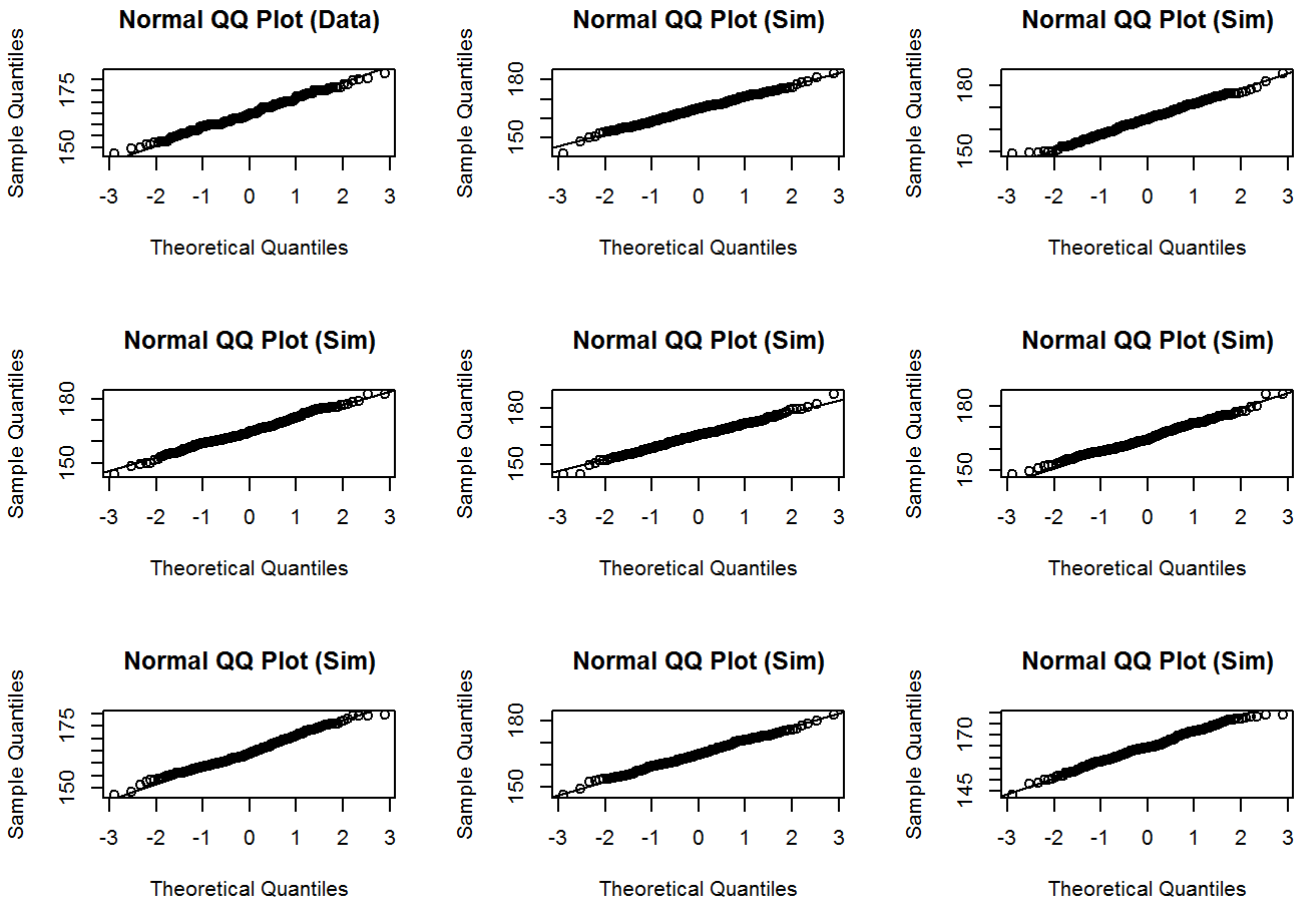
```
sim_norm <- rnorm(n = length(fdims$hgt), mean = fhgtmean, sd = fhgtsd)
qqnorm(fdims$hgt, main = "Actual Data")
qqline(fdims$hgt)
```




```
qqnorm(sim_norm, main = "Simmulated Data")
qqline(sim_norm)
```




```
qqnormsim(fdims$hgt)
```



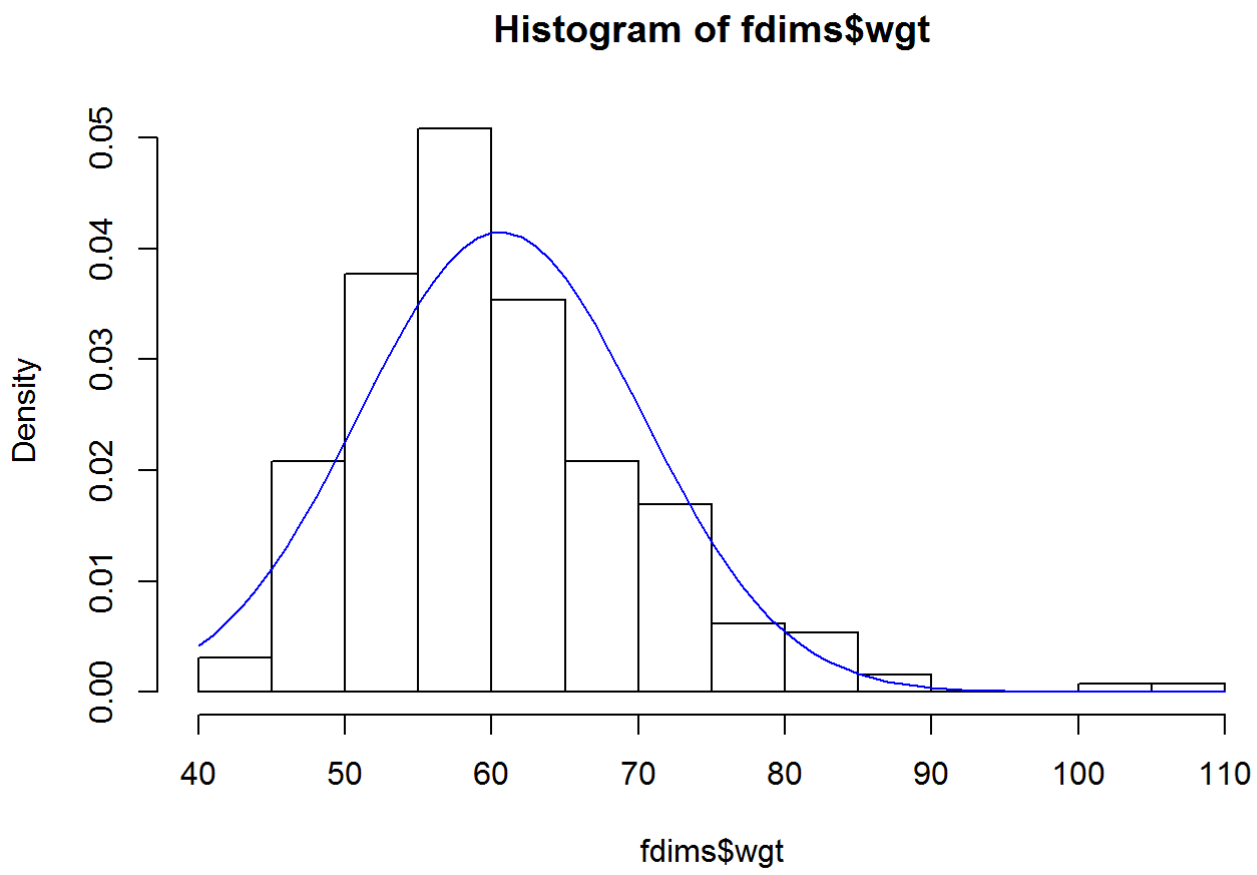
For the simmulated data, all most all points dall on the line and the line appear to be more smooth compare to that of actual data. However, even for the simmulated data, the tail is also appear to be distant to the normal probability line. The same patten is shown on the actual data.

Exercise 4: Does the normal probability plot for fdims$hgt look similar to the plots created for the simulated data? That is, do plots provide evidence that the female heights are nearly normal?

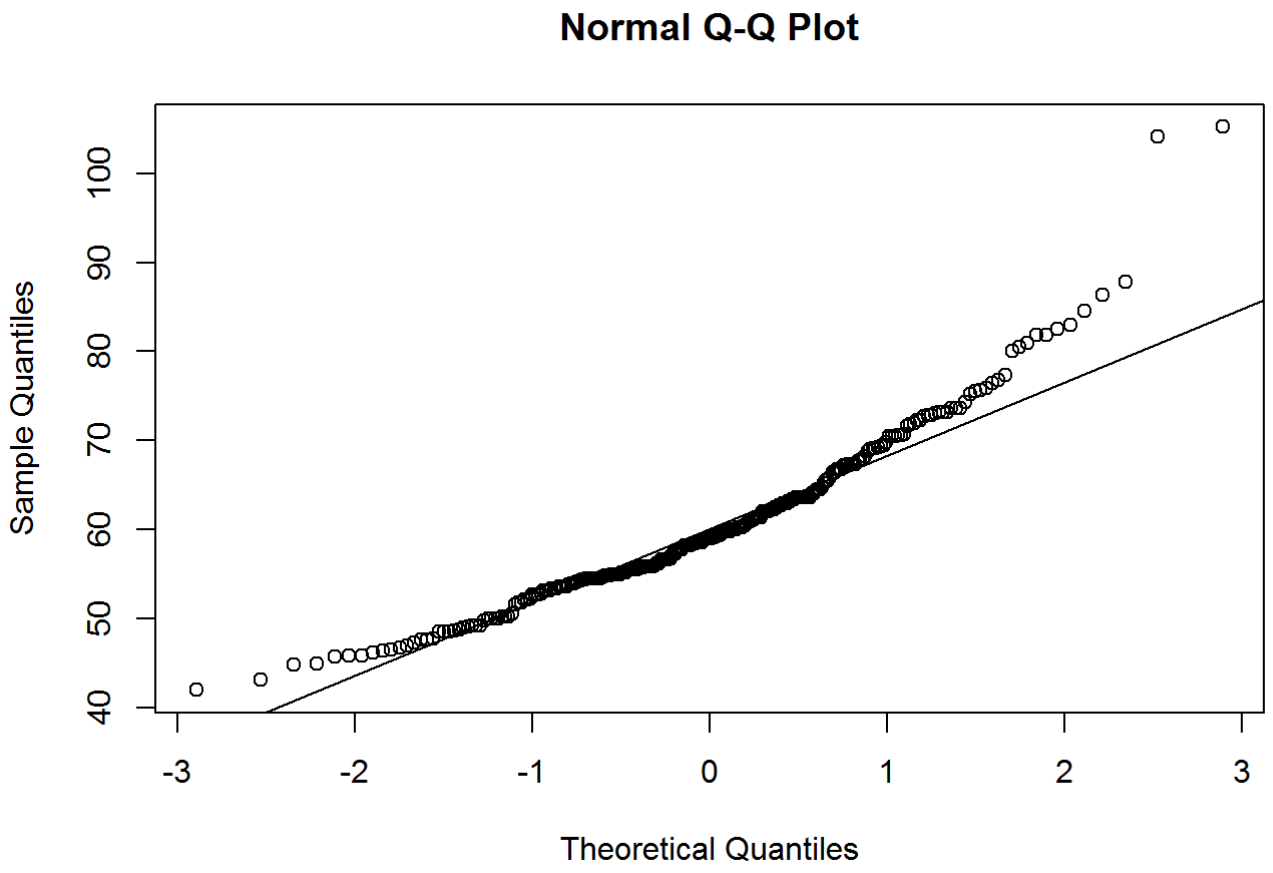
Yes, the normal probability plot for actual data is very similar to the simulated data. Therefore, female heights are nearly normally distributed.

Exercise 5: Using the same technique, determine whether or not female weights appear to come from a normal distribution.

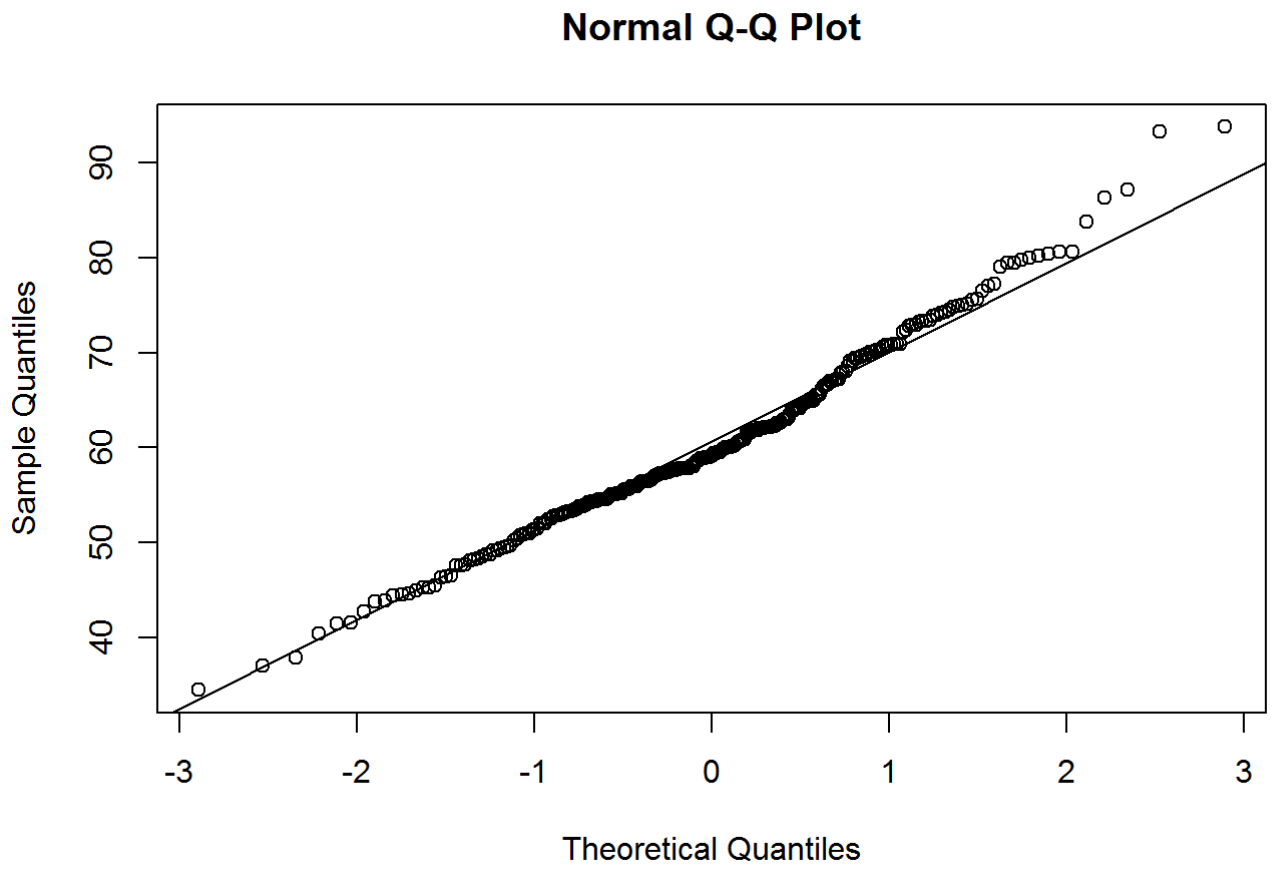
```
fwgtmean <- mean(fdims$wgt)
fwgtsd   <- sd(fdims$wgt)
hist(fdims$wgt, probability = TRUE)
x <- 40:110
y <- dnorm(x = x, mean = fwgtmean, sd = fwgtsd)
lines(x = x, y = y, col = "blue")
```



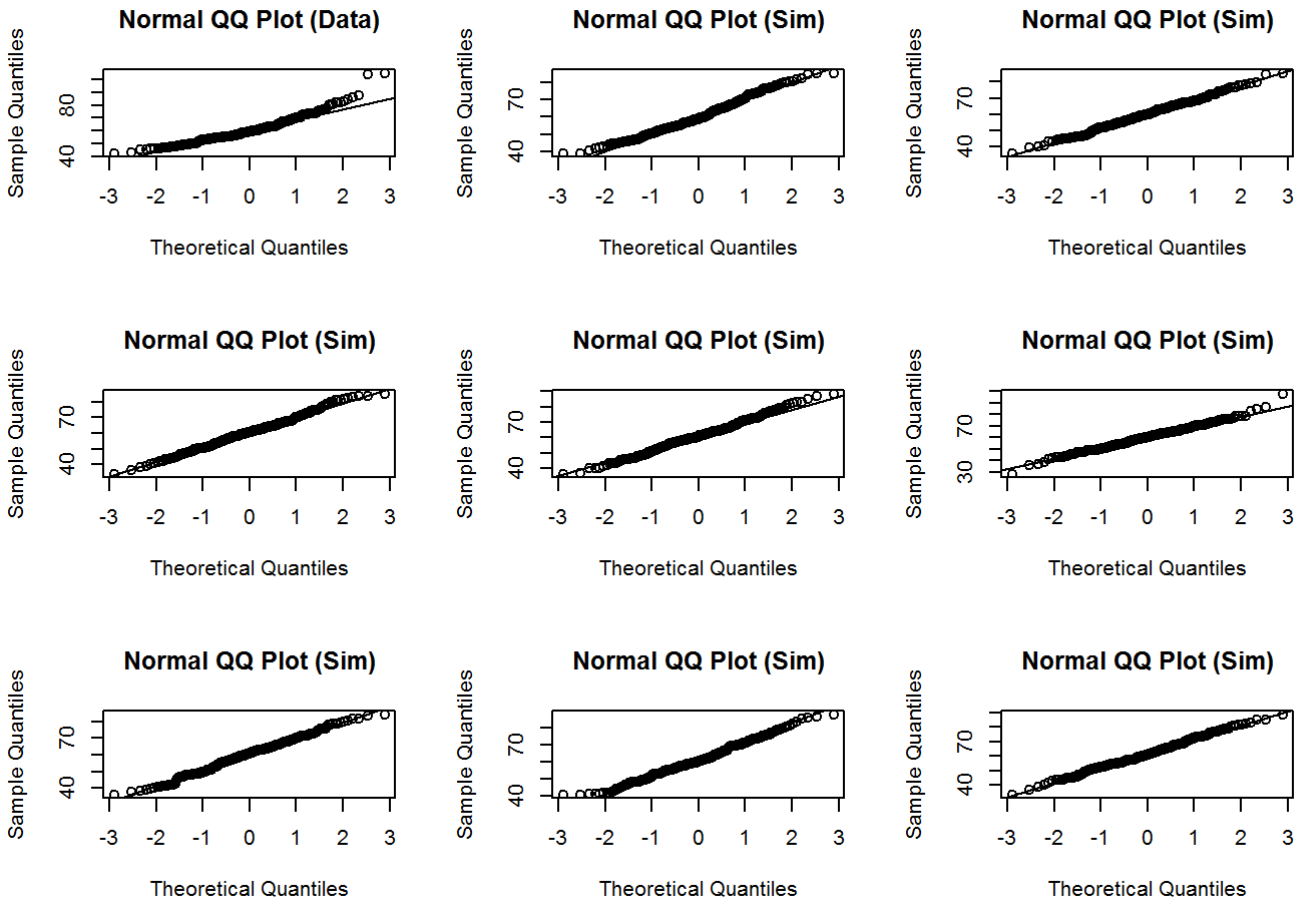

```
qqnorm(fdims$wgt)
qqline(fdims$wgt)
```




```
sim_norm <- rnorm(n = length(fdims$wgt), mean = fwgtmean, sd = fwgtsd)
qqnorm(sim_norm)
qqline(sim_norm)
```




```
qqnormsim(fdims$wgt)
```



Based on the those curves I got, we can tell the normal probability plot for the actual weight data is pretty much linear and they falls mostly on the normal distribution line. It resembles all other figures that are generated by theretical normal distributed data. However, the actual data is little bit skewed to teh right as the curve bend up on right side. In addition to that, two points on the top right side looks like they are completely off the tract.

Exercise 6: Write out two probability questions that you would like to answer; one regarding female heights and one regarding female weights. Calculate the those probabilities using both the theoretical normal distribution as well as the empirical distribution (four probabilities in all). Which variable, height or weight, had a closer agreement between the two methods?

A. What is the probability that female's height is above 180cm?

```
sim_norm <- rnorm(n = length(fdims$hgt), mean = fhgtmean, sd = fhgtsd)
1 - pnorm(q = 180, mean = fhgtmean, sd = fhgtsd)
```

```
## [1] 0.01040328
```

```
sum(fdims$hgt > 180) / length(fdims$hgt)
```

```
## [1] 0.007692308
```

These two probabilities are quite closer to each other. Height has a much closer agreement between the two methods.

B . What is the probability that female's weight is above 100kg?

```
sim_norm <- rnorm(n = length(fdims$wgt), mean = fwgtmean, sd = fwgtsd)
1 - pnorm(q = 100, mean = fwgtmean, sd = fwgtsd)
```

```
## [1] 2.088847e-05
```

```
sum(fdims$wgt > 100) / length(fdims$wgt)
```

```
## [1] 0.007692308
```

The theretical probability is much higher than the actual probability.

On Your Own Now let's consider some of the other variables in the body dimensions data set. Using the figures at the end of the exercises, match the histogram to its normal probability plot. All of the variables have been standardized (first subtract the mean, then divide by the standard deviation), so the units won't be of any help. If you are uncertain based on these figures, generate the plots in R to check.

a. The histogram for female biiliac (pelvic) diameter (bii.di) belongs to normal probability plot letter B.

b. The histogram for female elbow diameter (elb.di) belongs to normal probability plot letter C.

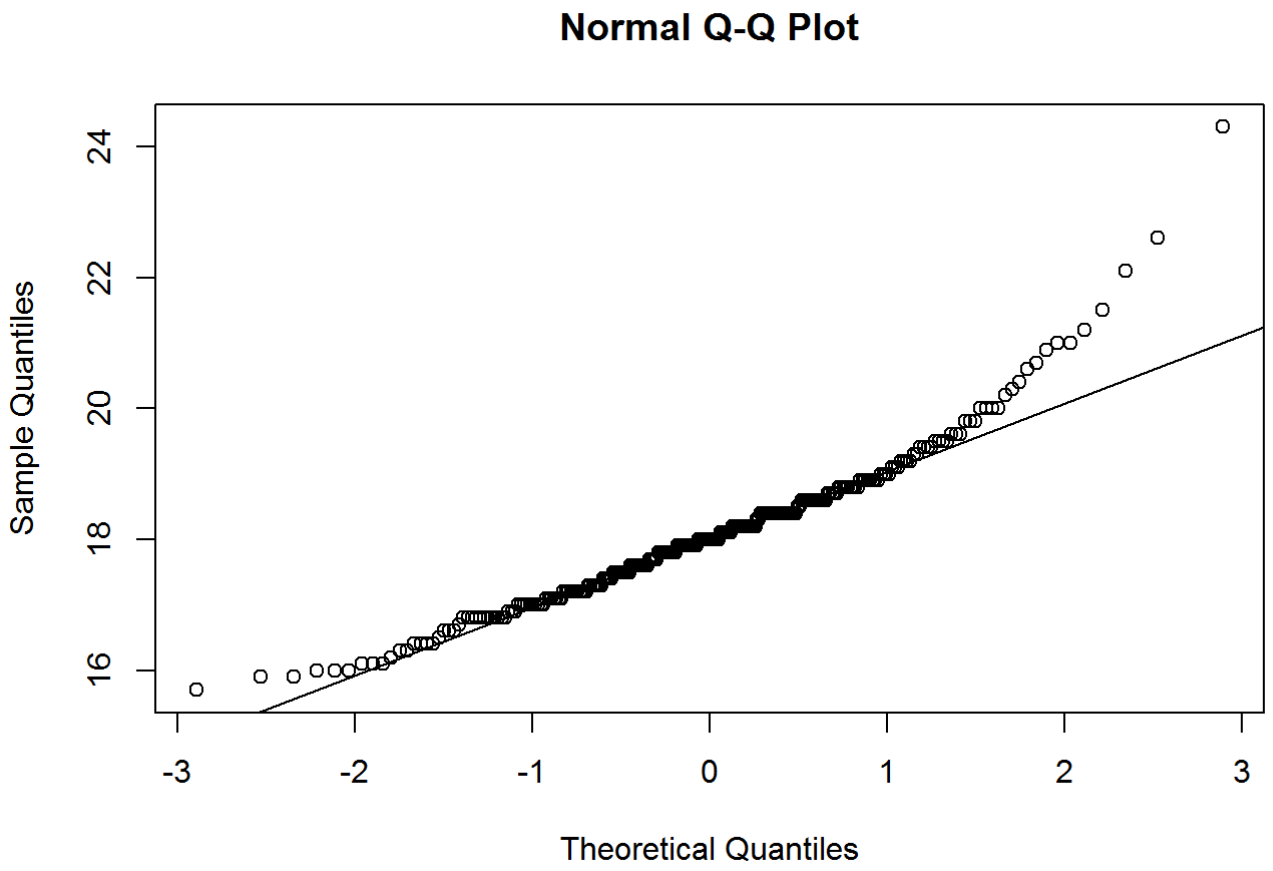
c. The histogram for general age (age) belongs to normal probability plot letter D.

d. The histogram for female chest depth (che.de) belongs to normal probability plot letter A.

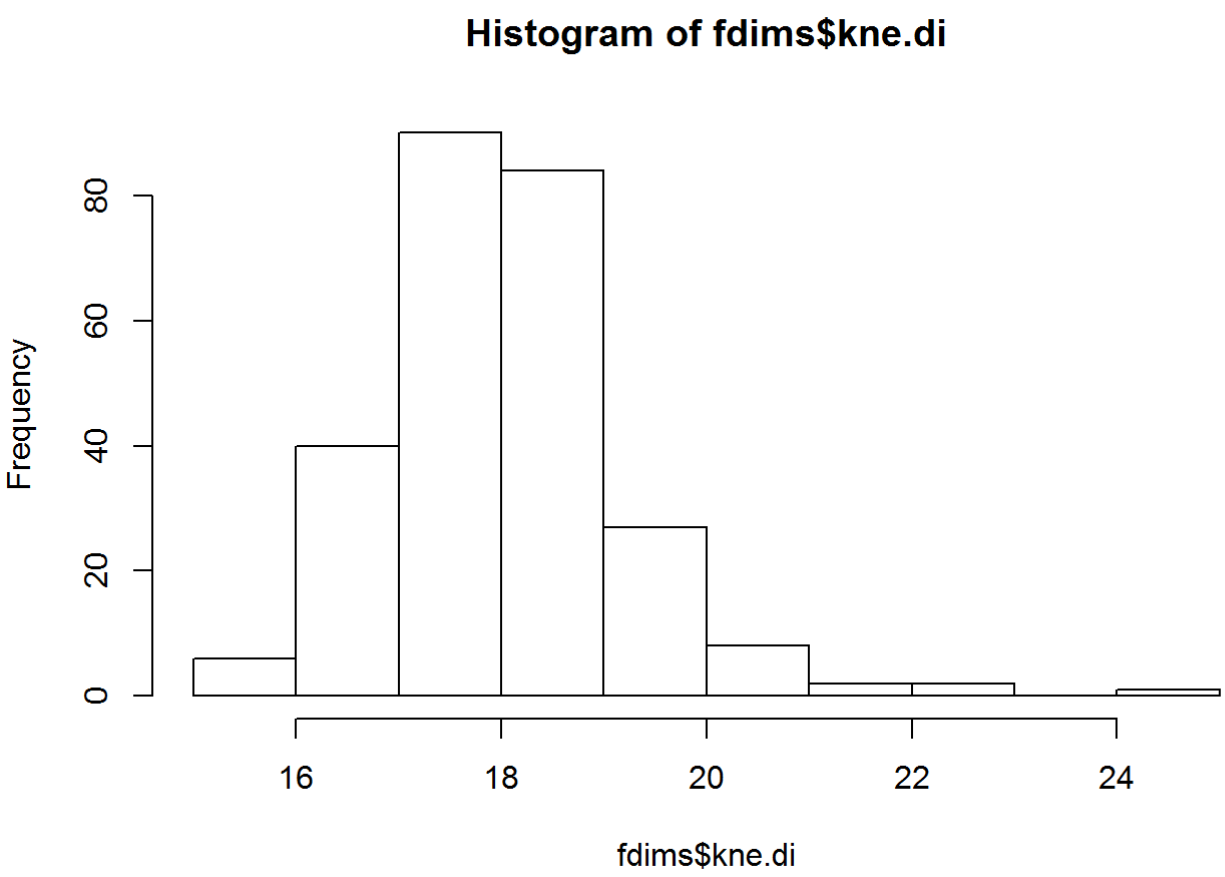
Note that normal probability plots C and D have a slight stepwise pattern. Why do you think this is the case? It is likely due to the fact a lots of those measurement get round up. For example, for age most people will report integers rather than some numbers with decimal points.

As you can see, normal probability plots can be used both to assess normality and visualize skewness. Make 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? Use a histogram to confirm your findings.

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




```
hist(fdims$kne.di)
```



The normal probability plot is bend upward on the right side, therefore, it is right skewed. The histogram proves this patten.