DA 606 - Lab 3

Dan Fanelli

February 24, 2016

Setup 1:

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

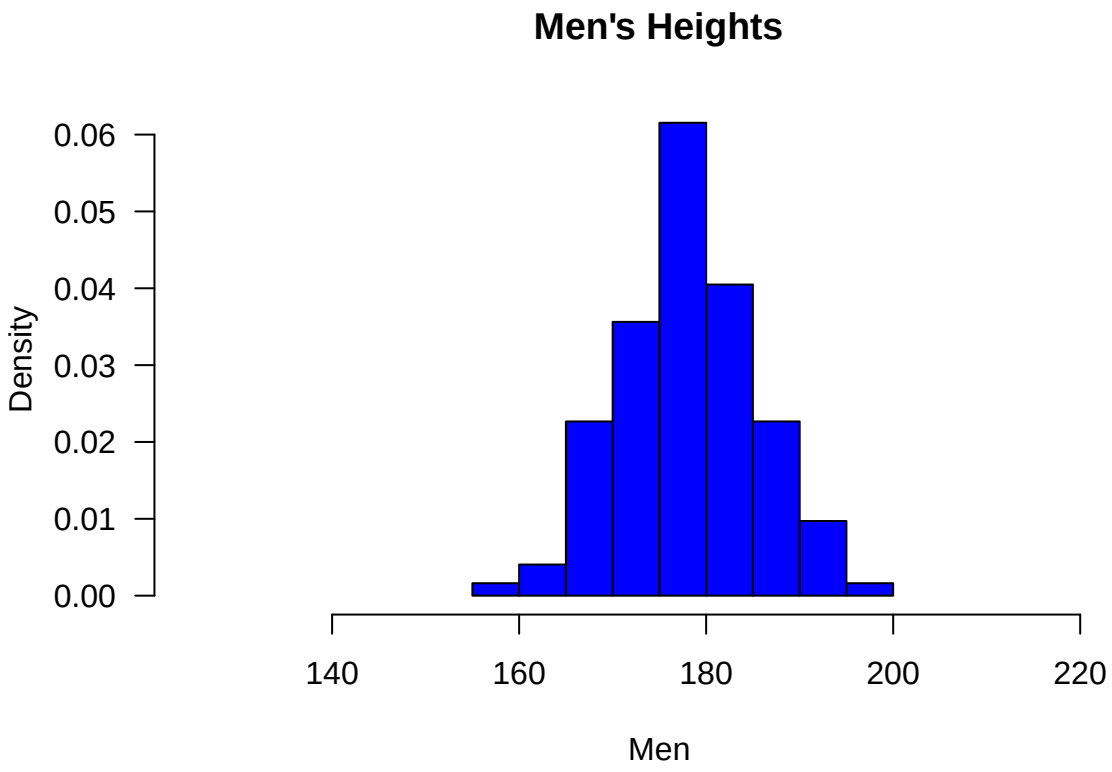
```
##   bia.di bii.di bit.di che.de che.di elb.di wri.di kne.di ank.di sho.gi
## 1   42.9   26.0   31.5   17.7   28.0   13.1   10.4   18.8   14.1  106.2
## 2   43.7   28.5   33.5   16.9   30.8   14.0   11.8   20.6   15.1  110.5
## 3   40.1   28.2   33.3   20.9   31.7   13.9   10.9   19.7   14.1  115.1
## 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)
```

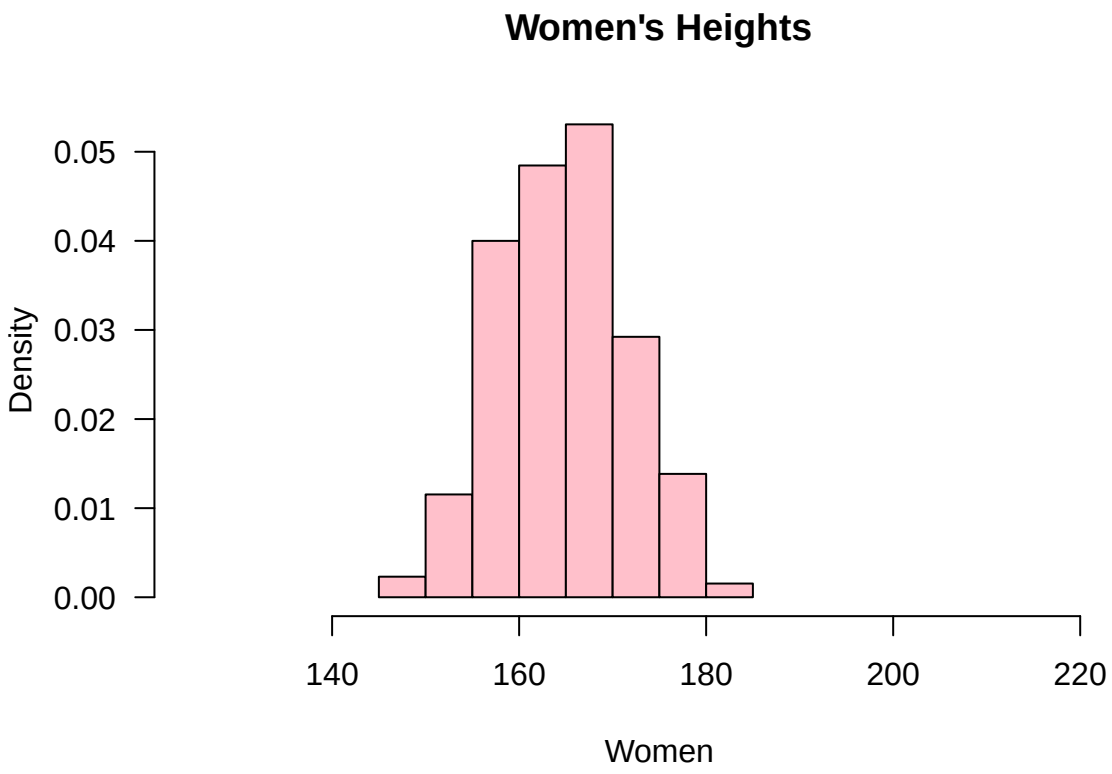
Exercise 1: The mean of the men's histogram is obviously further right, but it seems that the Men's is (slightly) skewed right while the women's is skewed left

```
hist(mdims$hgt,
     main="Men's Heights",
     xlab="Men",
     border="black",
     col="blue",
```

```
     xlim=c(125,225),
     las=1,
     breaks=10,
     prob = TRUE)
```

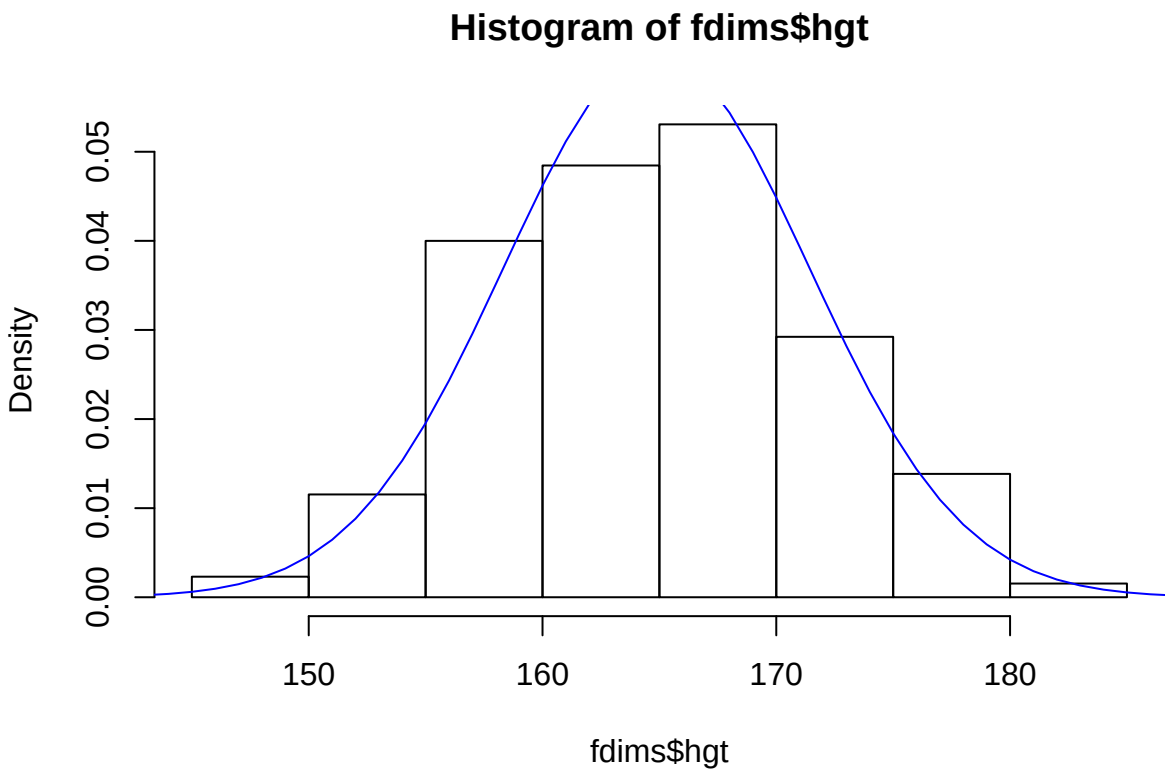



```
hist(fdims$hgt,
     main="Women's Heights",
     xlab="Women",
     border="black",
     col="pink",
     xlim=c(125,225),
     las=1,
     breaks=10,
     prob = TRUE)
```



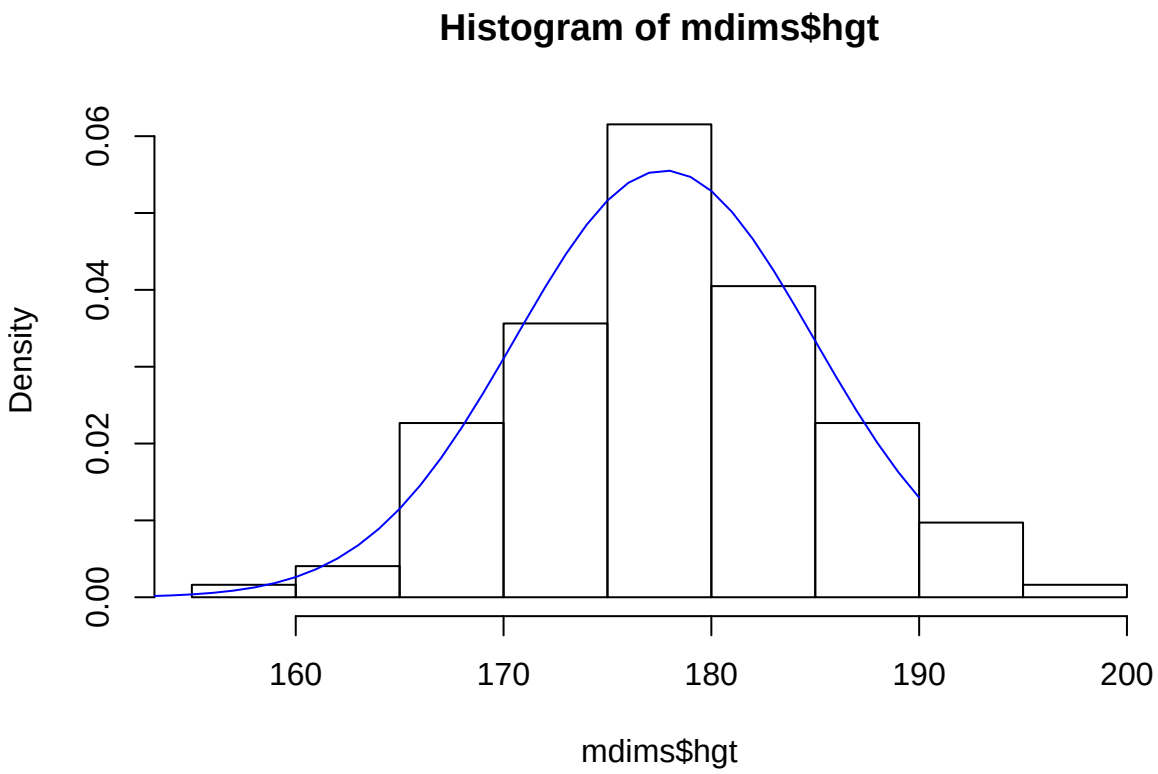
Setup 2: The given data for Females

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



Setup 2b: (Do it for Males too)

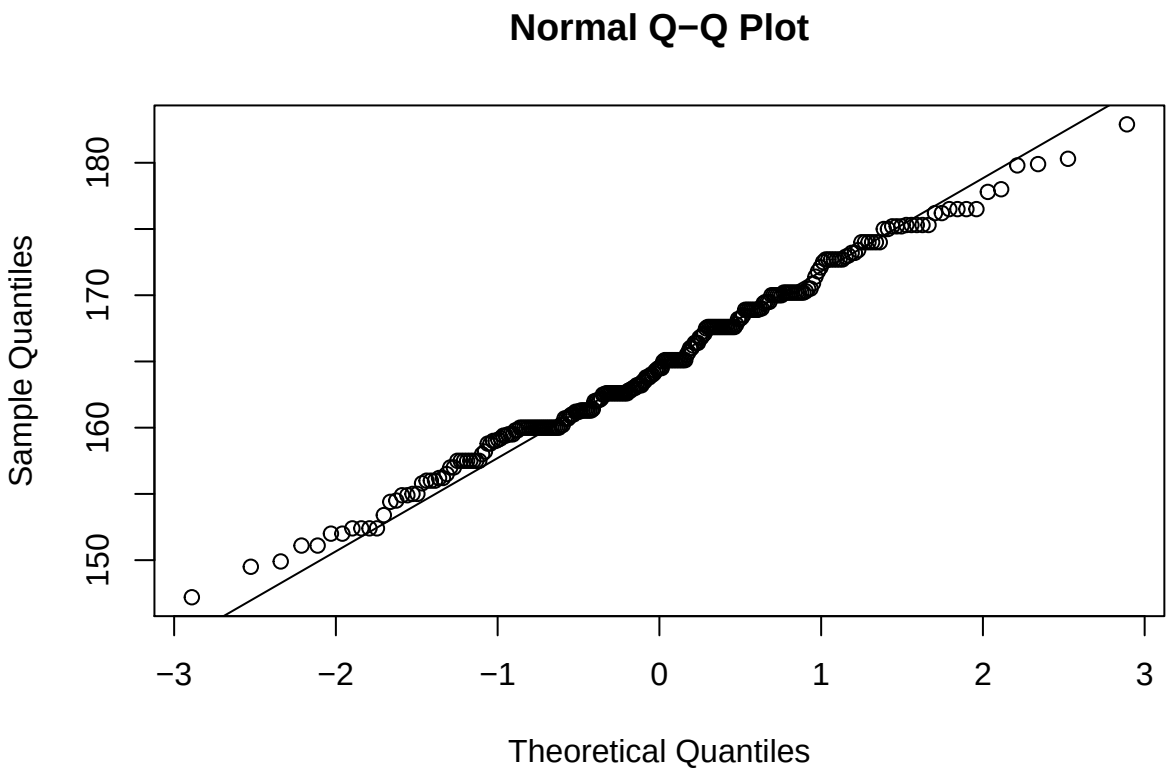
```
mhgtmean <- mean(mdims$hgt)
mhgtsd   <- sd(mdims$hgt)
hist(mdims$hgt, probability = TRUE)
x <- 140:190
y <- dnorm(x = x, mean = mhgtmean, sd = mhgtsd)
lines(x = x, y = y, col = "blue")
```



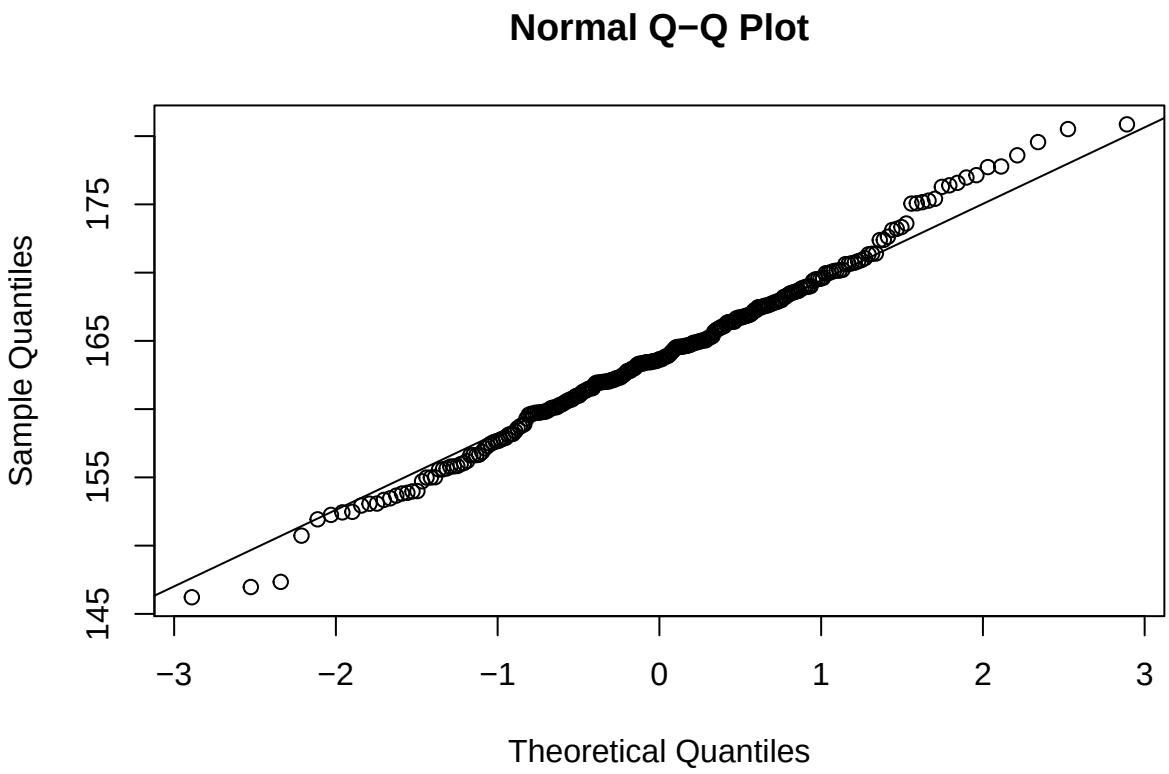
Exercise 2: Based on the data, and after looking at the men's chart, I would say that this chart does follow a nearly normal distribution, though I find it interesting that the female's curve does not go above the histogram bars, while the men's curve does...

Setup 3:

```
qqnorm(fdims$hgt)
qqline(fdims$hgt)
```



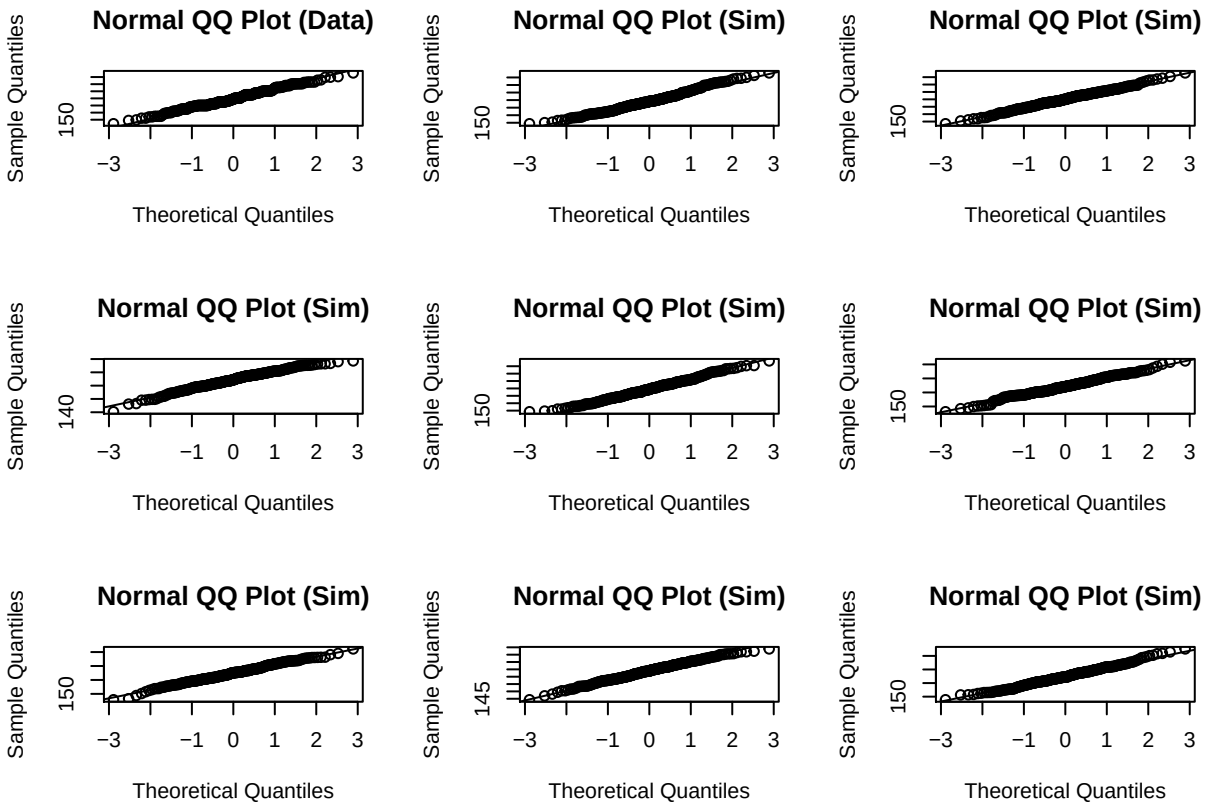

```
sim_norm <- rnorm(n = length(fdims$hgt), mean = fhgtmean, sd = fhgtsd)
qqnorm(sim_norm)
qqline(sim_norm)
```



Exercise 3: (See Plot Above) The simlated normal data looks comparable to the real fdim data.

Setup 4:

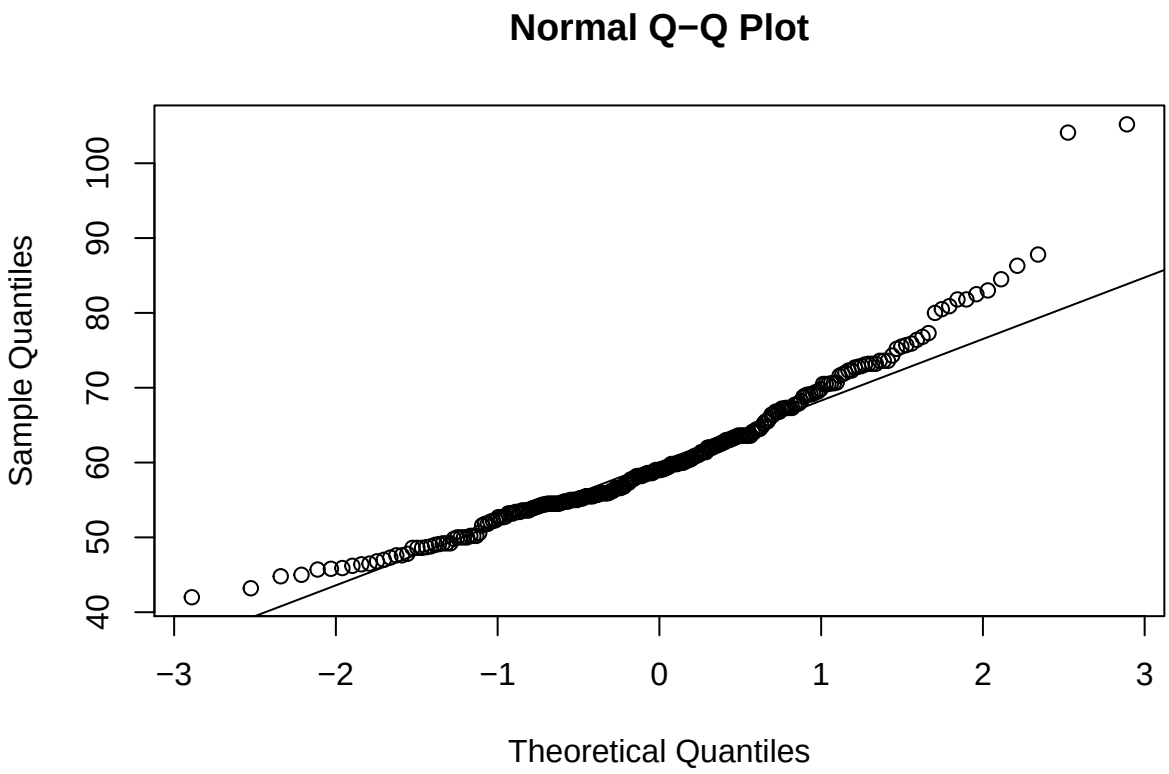
```
qqnormsim(fdims$hgt)
```



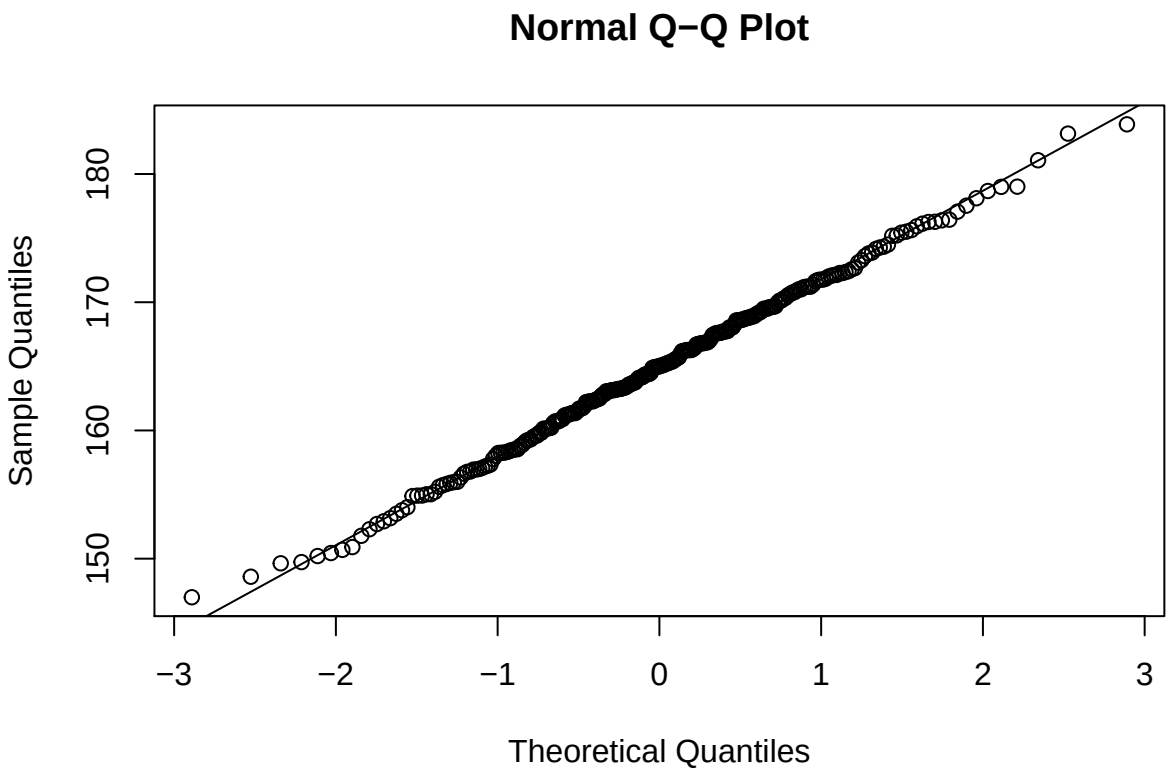
Exercise 4: The normal probability plot for fdims$hgt DOES look similar to the plots created for the simulated data. The plots DO provide evidence that the female heights are nearly normal.

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

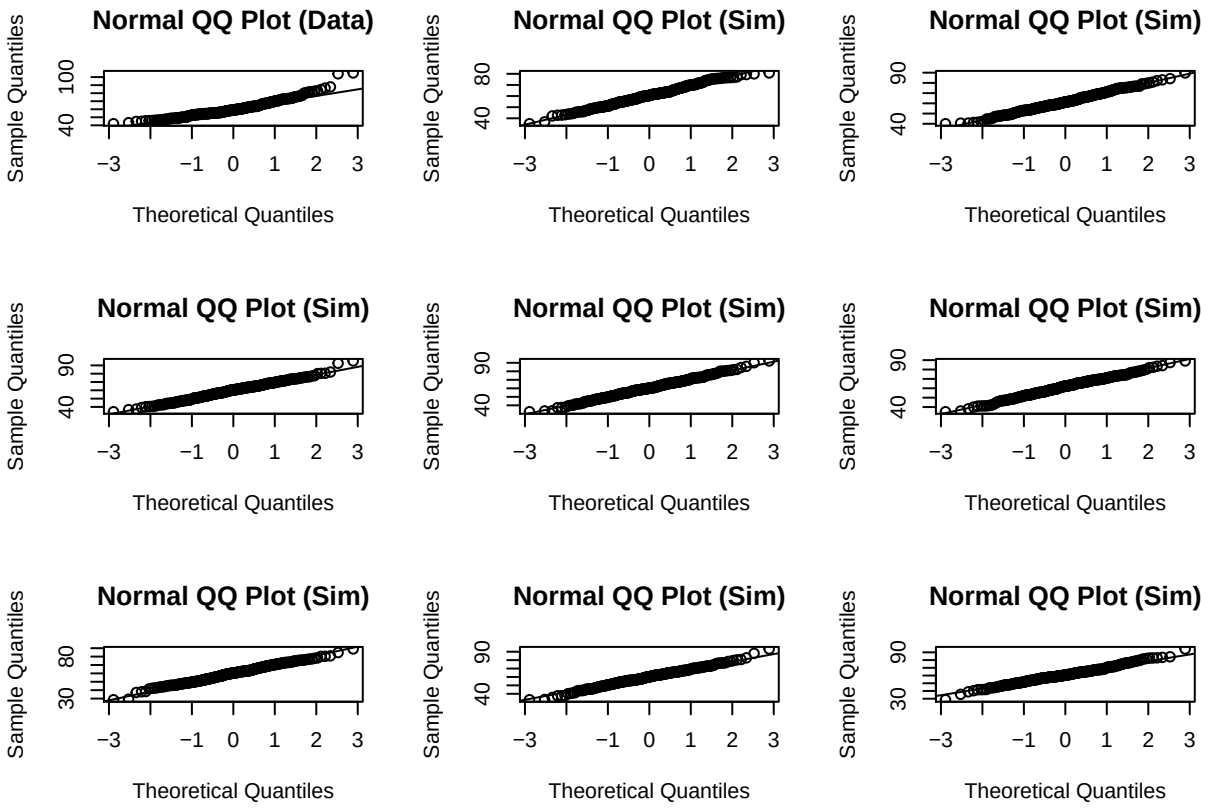
```
fwgtmean <- mean(fdims$wgt)
fwgtsd   <- sd(fdims$wgt)

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




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




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



Exercise 5: Aside from some outliers, the female weights DOES also appear to come from a normal distribution.

Setup 6:

```
1 - pnorm(q = 182, mean = fhgtmean, sd = fhgtsd)
```

```
## [1] 0.004434387
```

```
sum(fdims$hgt > 182) / length(fdims$hgt)
```

```
## [1] 0.003846154
```

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?

1: Female Heights What female height separates the highest 2.5% from the lowest 97.5%? **2: Female Weights** What female weight marks the highest 2.5% from the lowest 97.5%?

1 Heights as theoretical normal dist:

```
pop_sd <- sd(fdims$hgt)*sqrt((length(fdims$hgt)-1)/(length(fdims$hgt)))
pop_mean <- mean(fdims$hgt)
# 2 is the # of SD's for the 5%
value_for_97point5th_percentile <- pop_mean + 2 * pop_sd
value_for_97point5th_percentile
```

```
## [1] 177.9363
```

1 Heights as empirical dist:

```
sorted_f_heights <- sort(fdims$hgt, decreasing = TRUE)
num_heights <- length(sorted_f_heights)
count_of_top_2_pt_5_pct_f_height <- num_heights * 0.025
sorted_f_heights[count_of_top_2_pt_5_pct_f_height]
```

```
## [1] 177.8
```

2 Weights as theoretical normal dist:

```
pop_sd <- sd(fdims$wgt)*sqrt((length(fdims$wgt)-1)/(length(fdims$wgt)))
pop_mean <- mean(fdims$wgt)
# 2 is the # of SD's for the 5%
value_for_97point5th_percentile <- pop_mean + 2 * pop_sd
value_for_97point5th_percentile
```

```
## [1] 79.79476
```

2 Weights as empirical dist:

```
sorted_f_weights <- sort(fdims$wgt, decreasing = TRUE)
num_weights <- length(sorted_f_weights)
count_of_top_2_pt_5_pct_f_weight <- num_weights * 0.025
sorted_f_weights[count_of_top_2_pt_5_pct_f_weight]
```

```
## [1] 83
```

The **Heights** had a closer agreement between the two methods.