Lab4: Normal Probability Distributions

Quentin Terry

October 28, 2018

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

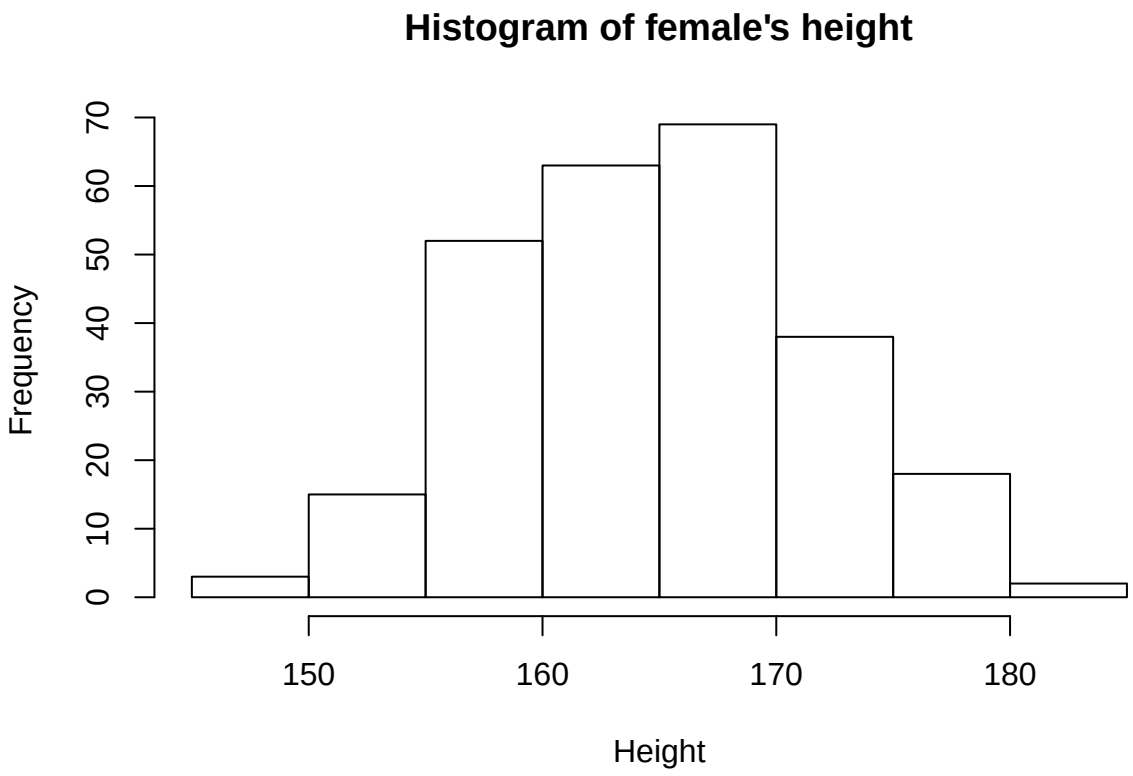
```
head(bdims)
```

```
##   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
```

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.)

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

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

hist(mdims$hgt, xlab = "Height", main = "Histogram of men's height")
```


```r
hist(fdims$hgt, xlab = "Height", main = "Histogram of female's height")
```




```
summary(mdims$hgt)
```

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

```
summary(fdims$hgt)
```

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

```
sd(mdims$hgt)
```

```
## [1] 7.183629
```

```
sd(fdims$hgt)
```

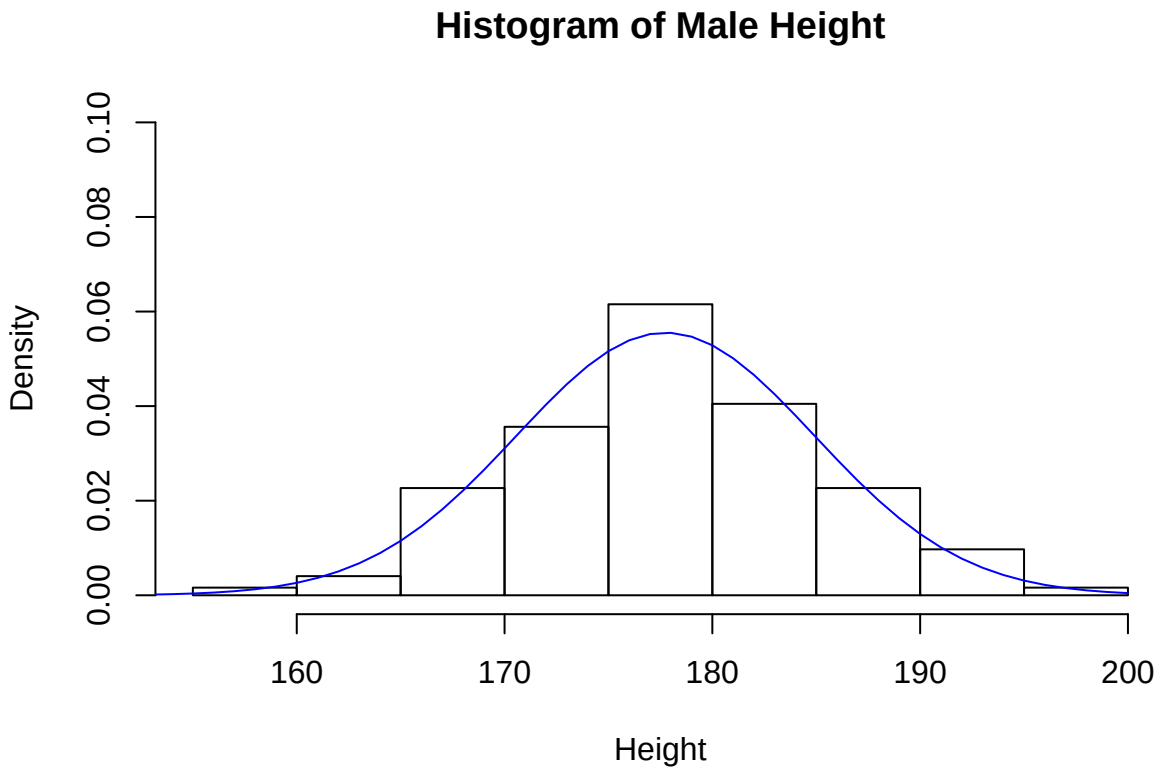
```
## [1] 6.544602
```

- The mens shape is more normal than females, however both male and females shapes are symetric. The center for the males is 177.8 whereas the females center is 164.5 meaning that the females heignt on average is less than the males height. Also the spread of the females height is 6.544602 which is less than the males at 7.183629 meaning that

```
mhgtmean <- mean(mdims$hgt)

mhgtsd <- sd(mdims$hgt)

mwgtmean <- mean(mdims$wgt)
```

```r
mwgtsd <- sd (mdims $wgt)
```

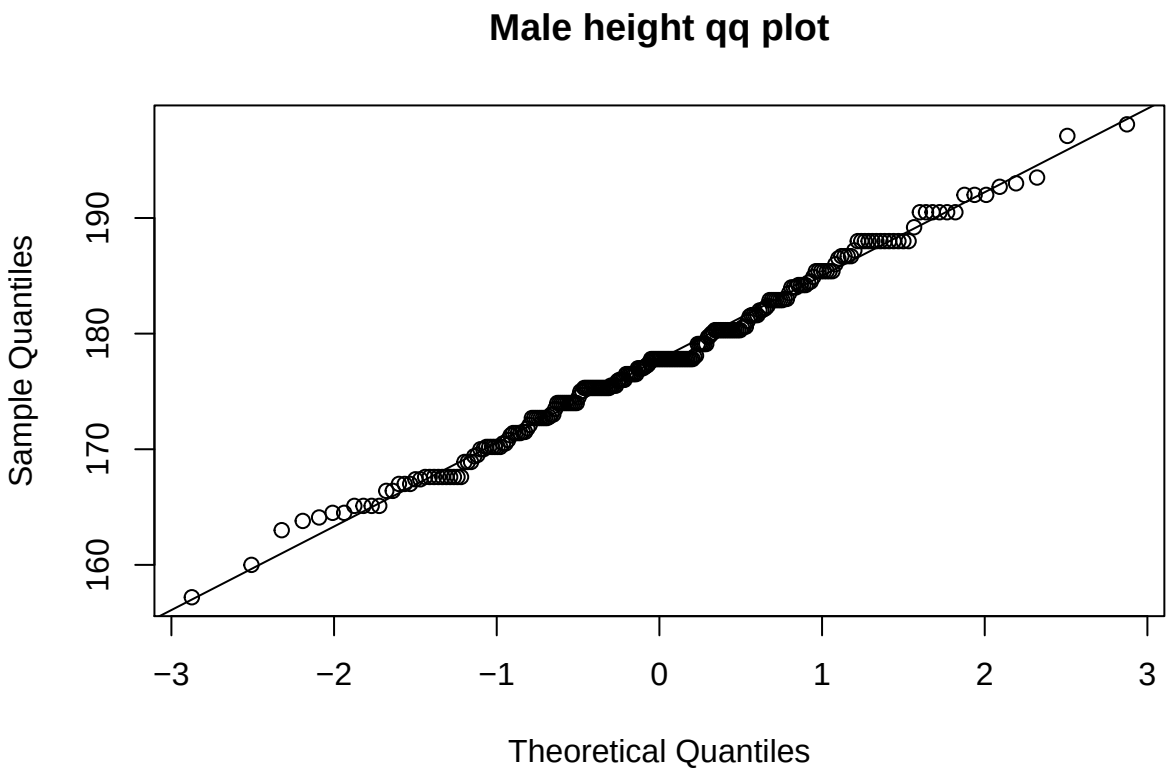
```r
hist(mdims$hgt, probability = TRUE, ylim = c(0,0.1), main = "Histogram of Male Height", xlab = "Height")

x <- 150:200

y <- dnorm(x = x, mean = mhgtmean, sd = mhgtsd)

lines(x = x, y = y, col = "blue")
```



Exercise 2: Based on this plot, does it appear that the men's height data follow a nearly normal distribution? Explain.

- Based on the plot, The histogram seems like a bell- shaped curve. Therefore, It apears that the men's height follows a nearly normal distribution.

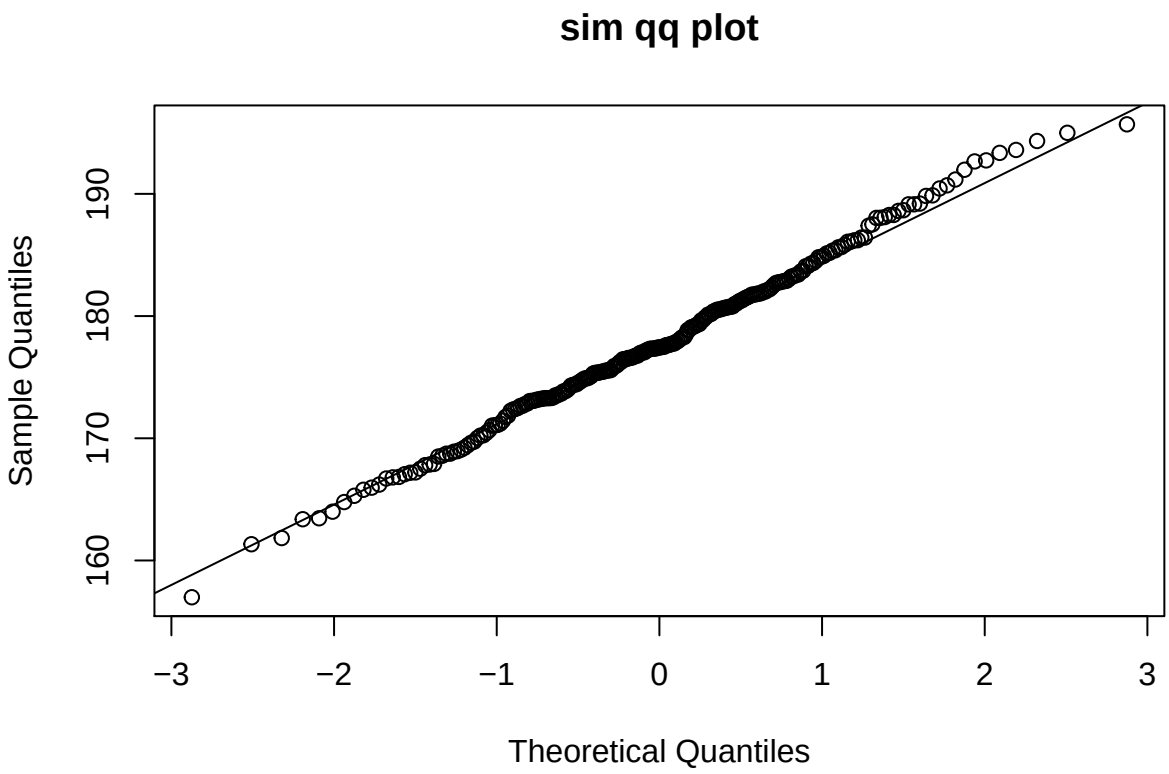
```r
qqnorm(mdims$hgt, main = "Male height qq plot")

qqline(mdims$hgt)
```




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

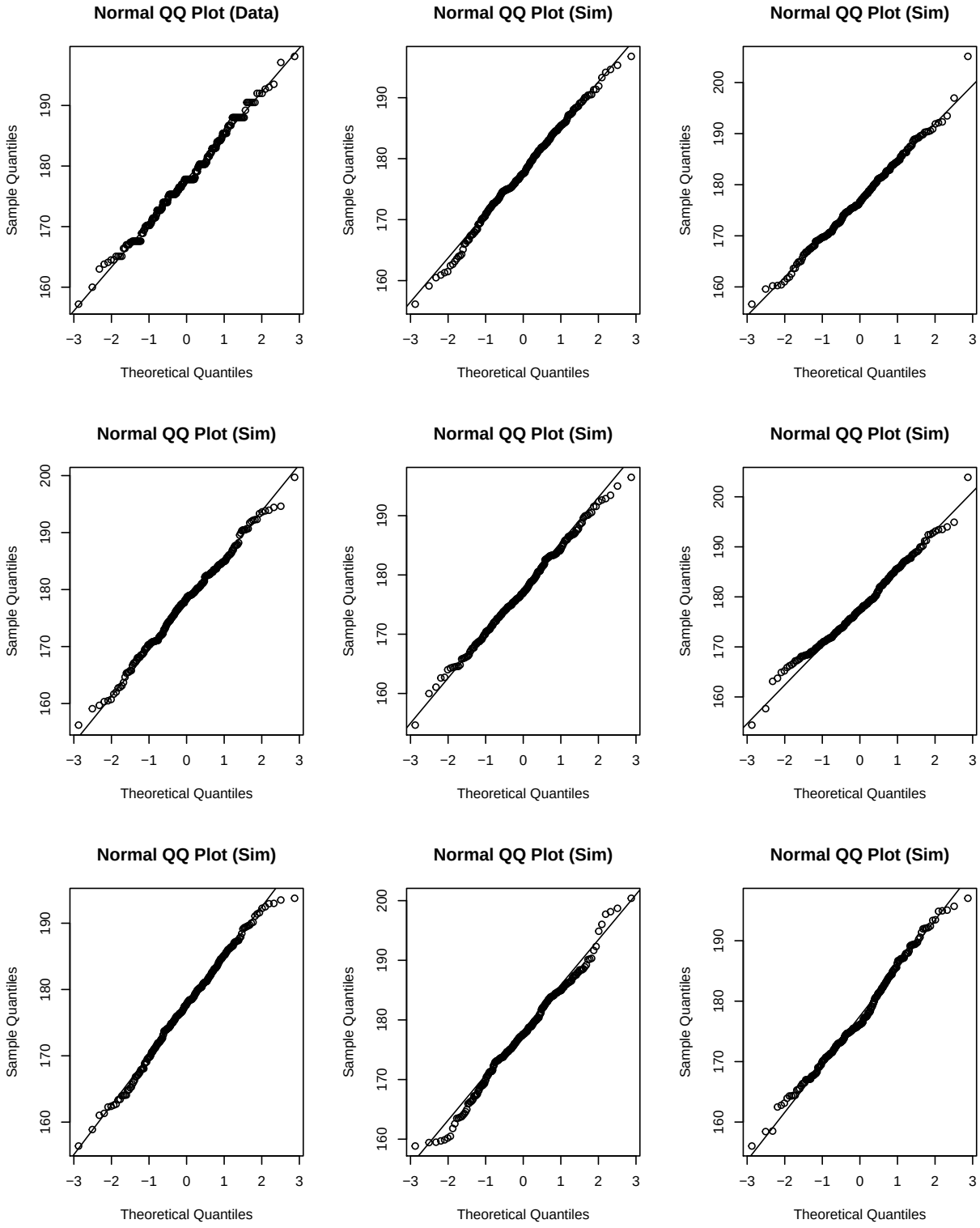
Exercise 3: Make a normal (Q-Q) probability plot of sim_norm. Do all of the points fall on the line? How does this plot compare to the normal probability plot for the actual data?

```
qqnorm(sim_norm, main = "sim qq plot")
qqline(sim_norm)
```



- All the points do not fall on the line but are close to the line. This plot is more linear than the actual data plot even though all the points of the simulated plot do not fall on the line.

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

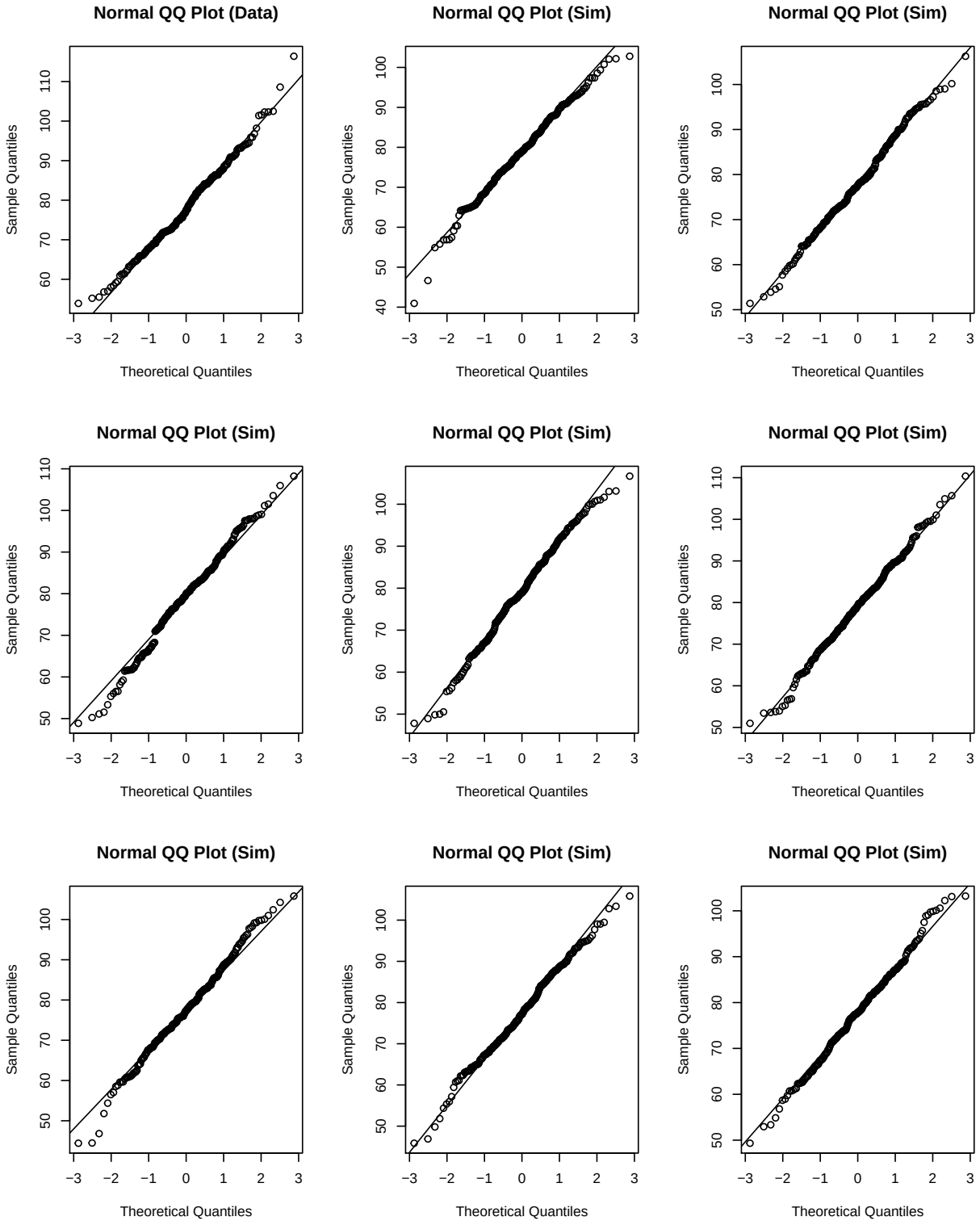

Normal QQ Plot (Data)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Sample Quantiles
Theoretical Quantiles

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.

- The normal plot for male heights is very similar to the simulated data. However, The original data has a stairstep shape to it.

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.

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



- Male weights apear to not come from a normal distribution due to the fact that due to the right tail being longer than what a normal distribution would look like.

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.

- What is the probablility that a random male has a height taller than 180.0 cm?

```
sum(mdims$hgt > 180.0)/length(mdims$hgt)
```

```
## [1] 0.3724696
```

```
 1- pnorm(180, mhgtmean, mhgtsd)
```

```
## [1] 0.3768136
```

- What is the probablility that a random male is heavier than 80.7 Pounds?

```
sum(mdims$wgt > 80.7)/length(mdims$wgt)
```

```
## [1] 0.417004
```

```
 1- pnorm(80.7, mwgtmean, mwgtsd)
```

```
## [1] 0.403972
```

- Height is closer to normal due to the fact that the valuesof empirical distribution and theoretical normal distribution with height are closer together.

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) The histogram for general (i.e., male and female) age (age) corresponds to normal probability plot letter . (2) The histogram for female chest depth (che.de) corresponds to normal probability plot letter . (3) The histogram for female biiliac (pelvic) diameter (bii.di) corresponds to normal probability plot letter . (4) The histogram for female elbow diameter (elb.di) corresponds to normal probability plot letter

- 1) D

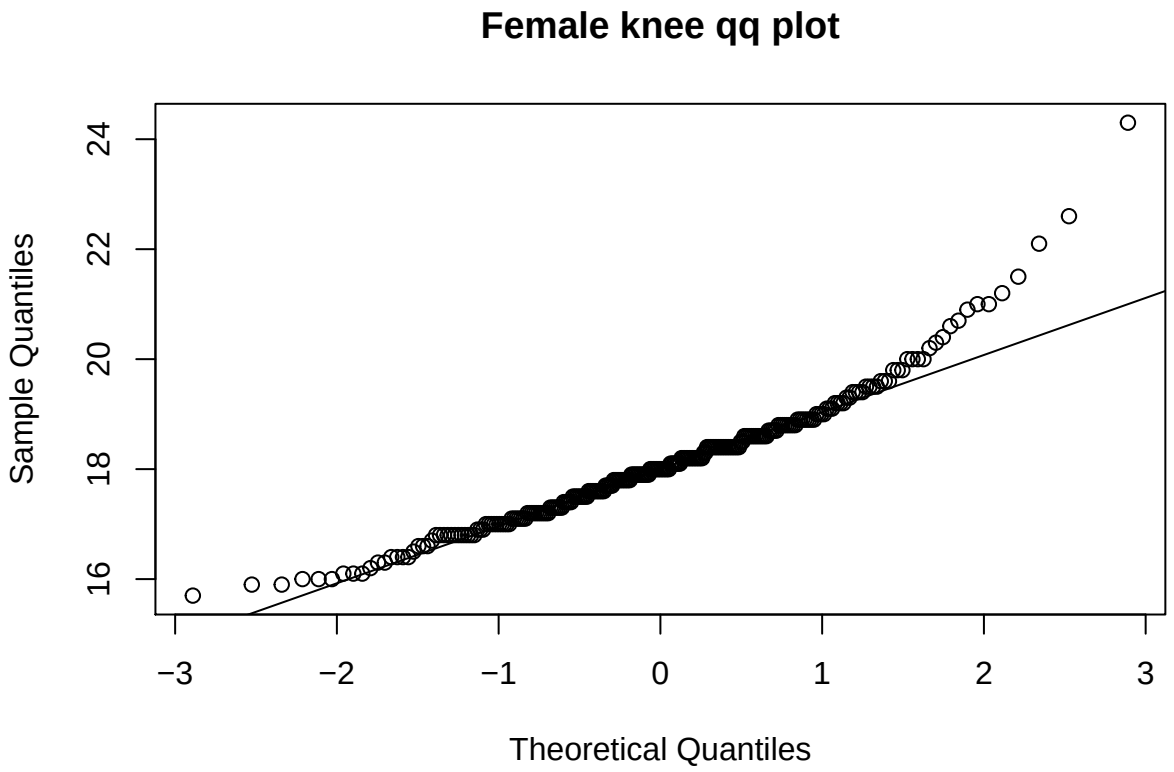
2) A
3) B
4) C

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

- This is due to the numbers used in the plot. because in the plot, age uses discrete values instead of continuous values, This creates the stepwise pattern.

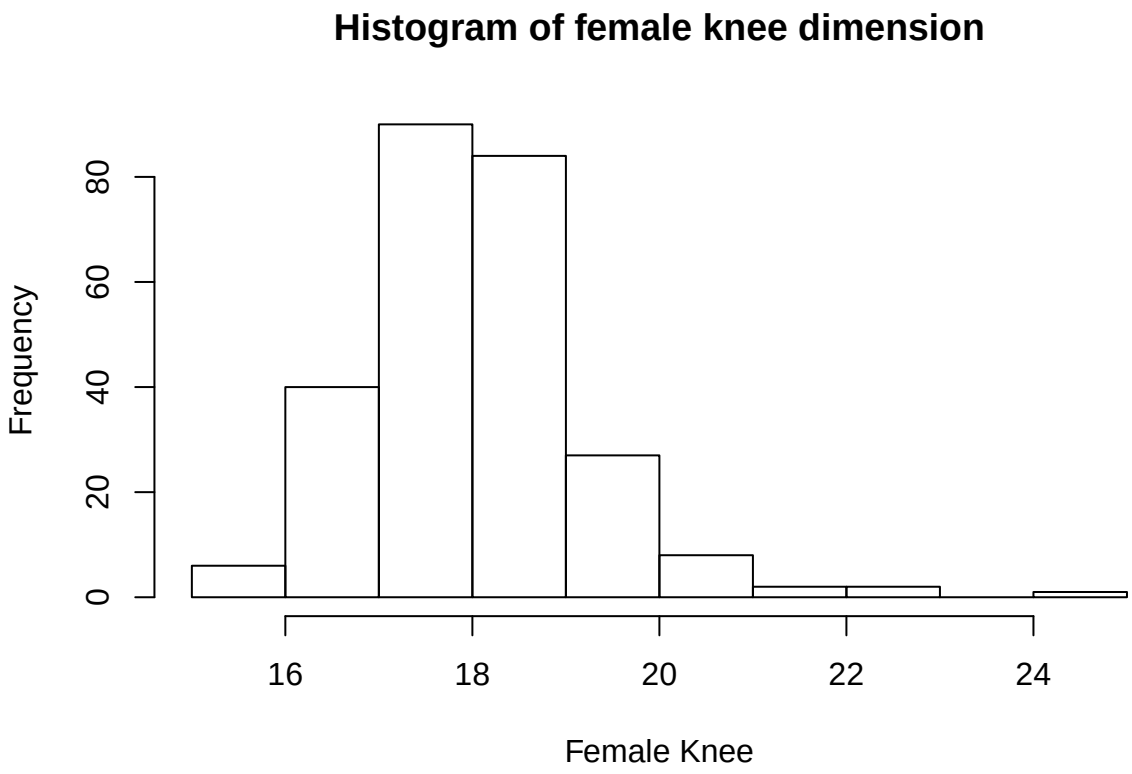
3. As you can see, normal probability plots can be used both to assess normality and visualize skewness. Produce a normal probability plot for female knee diameter (kne.di). Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Explain your reasoning. Use a histogram to confirm your findings.

```
qqnorm(fdims$kne.di, main = "Female knee qq plot")
qqline(fdims$kne.di)
```



- Based on the plot this variable is right skewed.

```
hist(fdims$kne.di, main = "Histogram of female knee dimension", xlab = "Female Knee")
```


Histogram of female knee dimension
Frequency
0
20
40
60
80
16
18
20
22
24
Female Knee