The normal distribution

In this lab we'll investigate the probability distribution that is most central to statistics: the normal distribution. If we are confident that our data are nearly normal, that opens the door to many powerful statistical methods. Here we'll use the graphical tools of R to assess the normality of our data and also learn how to generate random numbers from a normal distribution.

The Data

This week we'll be working with measurements of body dimensions. This data set contains measurements from 247 men and 260 women, most of whom were considered healthy young adults.

```
load("more/bdims.RData")
```

Let's take a quick peek at the first few rows of the data.

```
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
```

You'll see that for every observation we have 25 measurements, many of which are either diameters or girths. A key to the variable names can be found at http://www.openintro.org/stat/data/bdims.php, but we'll be focusing on just three columns to get started: weight in kg (`wgt`), height in cm (`hgt`), and `sex` (`1` indicates male, `0` indicates female).

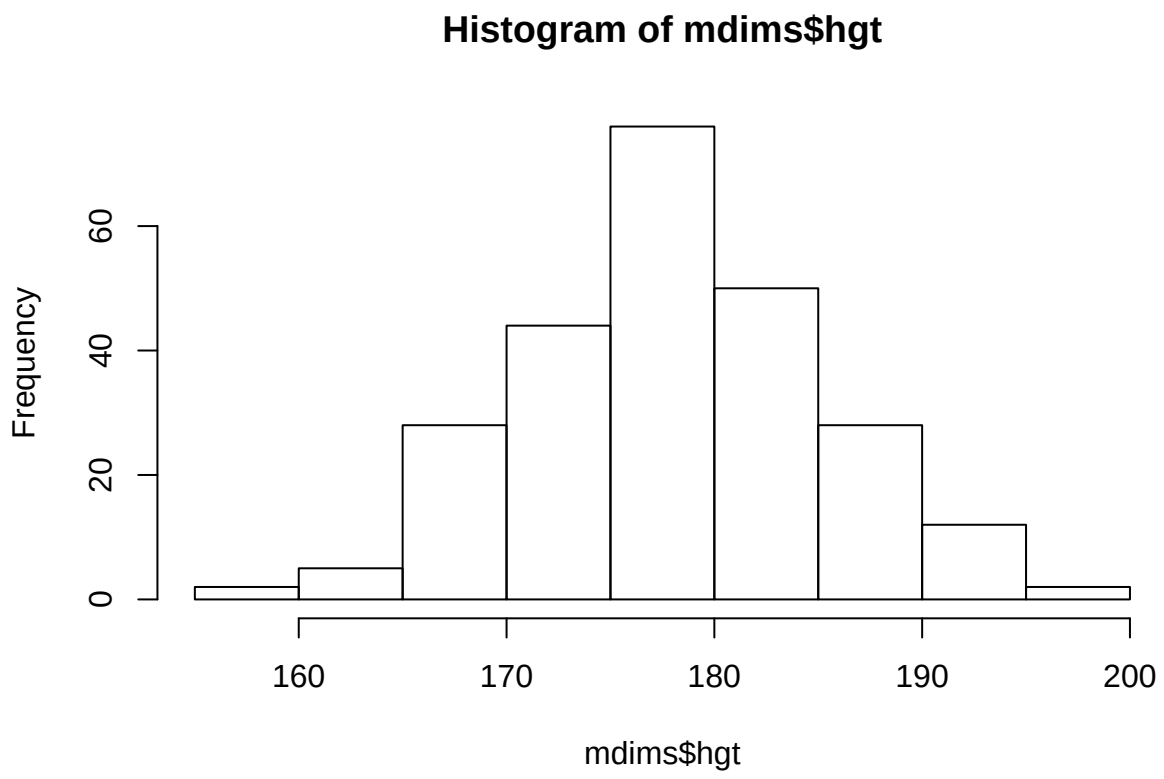
Since males and females tend to have different body dimensions, it will be useful to create two additional data sets: one with only men and another with only women.

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

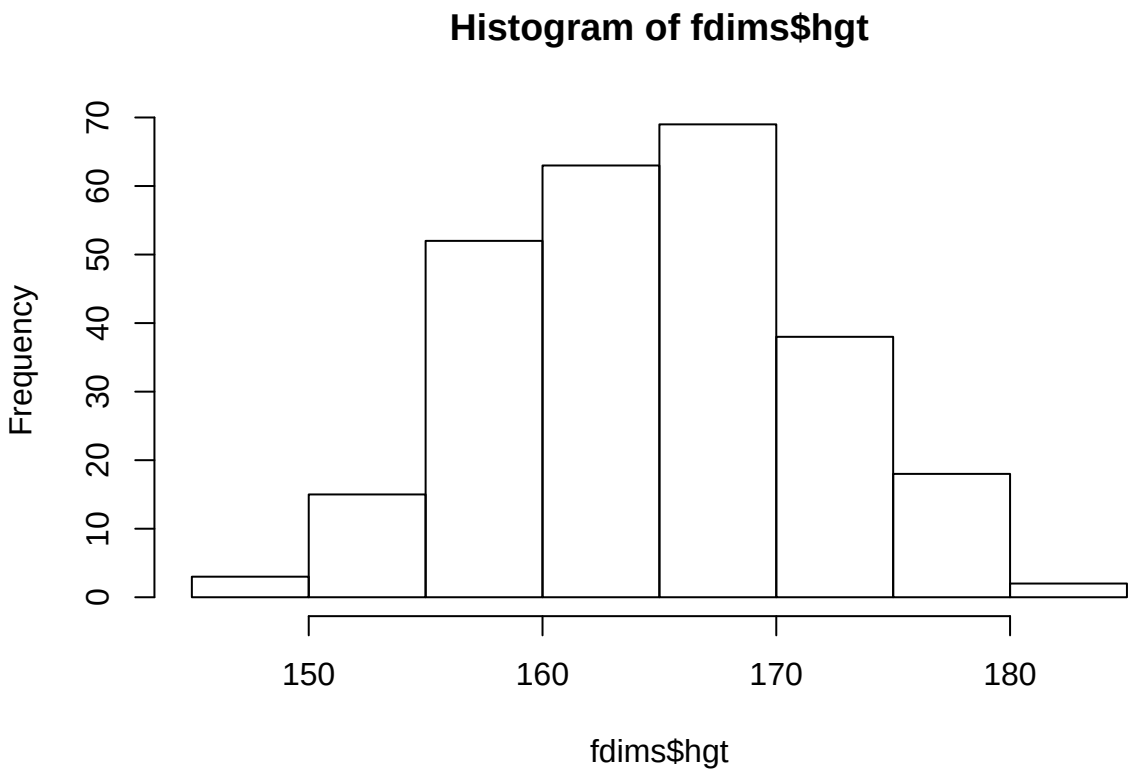
1. Make a histogram of men's heights and a histogram of women's heights. How would you compare the various aspects of the two distributions?

See below for my r code to generate the two histograms for men's heights and women's heights:

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




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



As you can see from the histograms, both seem to exhibit relatively normal distributions. The mean men's height is 177.75 centimeters (+/- 7.18). While the mean women's height is 164.87 (+/- 6.54).

```
library(psych)

describe(mdims$hgt)
```

```
##    vars   n   mean   sd median trimmed  mad   min   max range skew
## X1    1 247 177.75 7.18  177.8  177.67 7.41 157.2 198.1  40.9  0.1
##    kurtosis   se
## X1    -0.16 0.46
```

```
describe(fdims$hgt)
```

```
##    vars   n   mean   sd median trimmed  mad   min   max range skew
## X1    1 260 164.87 6.54  164.5  164.84 6.67 147.2 182.9  35.7 0.07
##    kurtosis   se
## X1    -0.32 0.41
```

The normal distribution

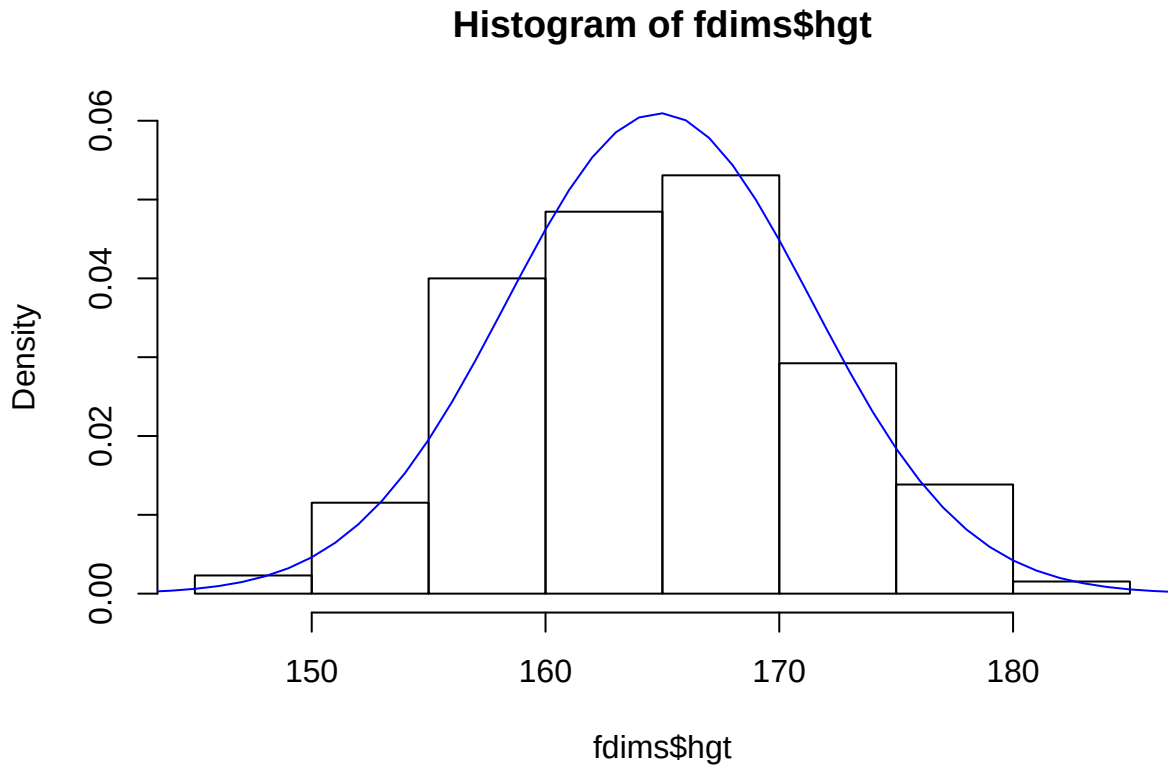
In your description of the distributions, did you use words like *bell-shaped* or *normal*? It's tempting to say so when faced with a unimodal symmetric distribution.

To see how accurate that description is, we can plot a normal distribution curve on top of a histogram to see how closely the data follow a normal distribution. This normal curve should have the same mean and standard deviation as the data. We'll be working with women's heights, so let's store them as a separate object and then calculate some statistics that will be referenced later.

```
fhgtmean <- mean(fdims$hgt)
fhgtsd   <- sd(fdims$hgt)
```

Next we make a density histogram to use as the backdrop and use the `lines` function to overlay a normal probability curve. The difference between a frequency histogram and a density histogram is that while in a frequency histogram the *heights* of the bars add up to the total number of observations, in a density histogram the *areas* of the bars add up to 1. The area of each bar can be calculated as simply the height *times* the width of the bar. Using a density histogram allows us to properly overlay a normal distribution curve over the histogram since the curve is a normal probability density function. Frequency and density histograms both display the same exact shape; they only differ in their y-axis. You can verify this by comparing the frequency histogram you constructed earlier and the density histogram created by the commands below.

```
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")
```



After plotting the density histogram with the first command, we create the x- and y-coordinates for the normal curve. We chose the `x` range as 140 to 190 in order to span the entire range of `fheight`. To create `y`, we use `dnorm` to calculate the density of each of those x-values in a distribution that is normal with mean `fhgtmean` and standard deviation `fhgtsd`. The final command draws a curve on the existing plot (the

density histogram) by connecting each of the points specified by `x` and `y`. The argument `col` simply sets the color for the line to be drawn. If we left it out, the line would be drawn in black.

The top of the curve is cut off because the limits of the x- and y-axes are set to best fit the histogram. To adjust the y-axis you can add a third argument to the histogram function: `ylim = c(0, 0.06)`.

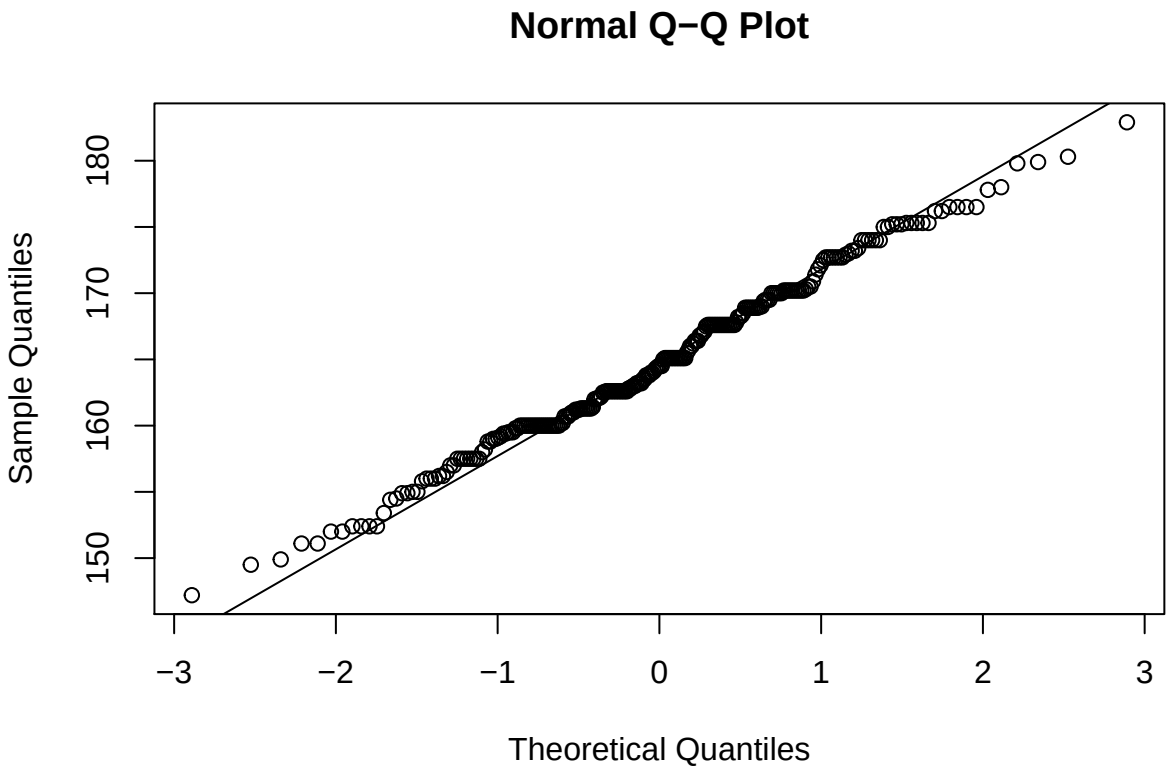
2. Based on the this plot, does it appear that the data follow a nearly normal distribution?

Based on the plot above it appears that the data is slightly skewed relative to a normal distribution (slightly left skewed), however it appears to *nearly* follow the normal distribution.

Evaluating the normal distribution

Eyeballing the shape of the histogram is one way to determine if the data appear to be nearly normally distributed, but it can be frustrating to decide just how close the histogram is to the curve. An alternative approach involves constructing a normal probability plot, also called a normal Q-Q plot for "quantile-quantile".

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



A data set that is nearly normal will result in a probability plot where the points closely follow the line. Any deviations from normality leads to deviations of these points from the line. The plot for female heights shows points that tend to follow the line but with some errant points towards the tails. We're left with the same problem that we encountered with the histogram above: how close is close enough?

A useful way to address this question is to rephrase it as: what do probability plots look like for data that I *know* came from a normal distribution? We can answer this by simulating data from a normal distribution using `rnorm`.

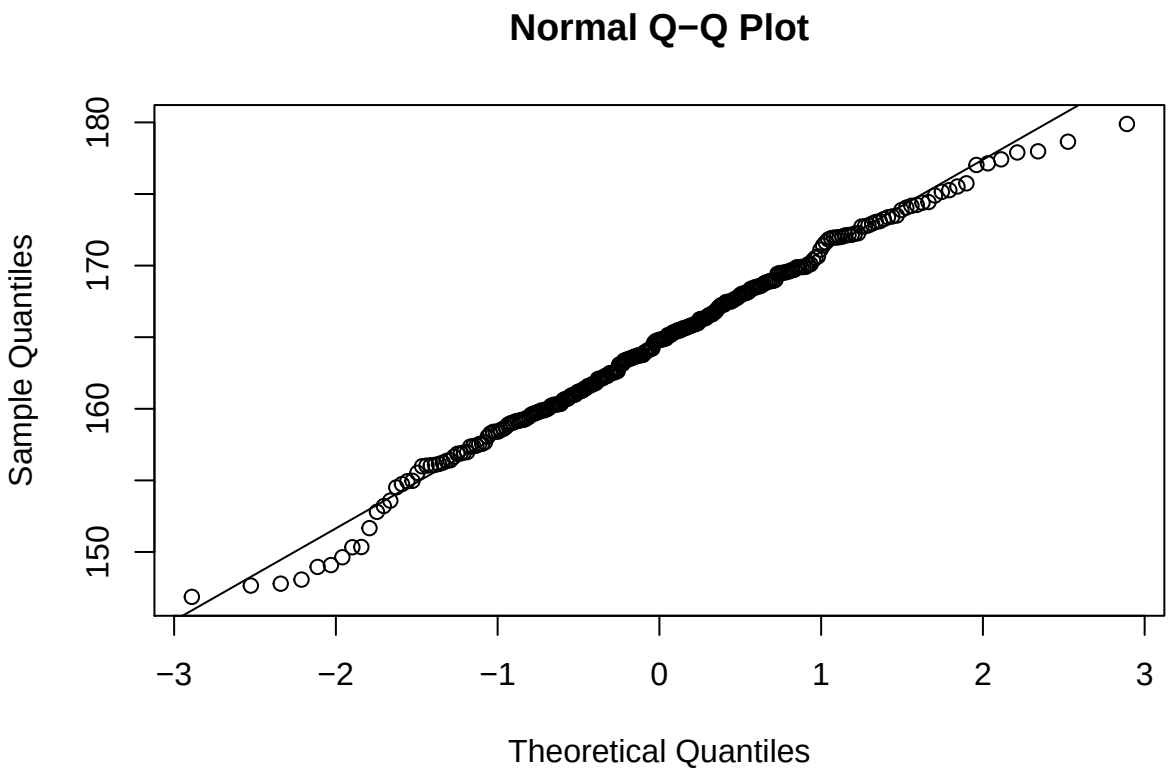
```
sim_norm <- rnorm(n = length(fdims$hgt), mean = fhgtmean, sd = fhgtsd)
```

The first argument indicates how many numbers you'd like to generate, which we specify to be the same number of heights in the `fdims` data set using the `length` function. The last two arguments determine the mean and standard deviation of the normal distribution from which the simulated sample will be generated. We can take a look at the shape of our simulated data set, `sim_norm`, as well as its normal probability plot.

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?

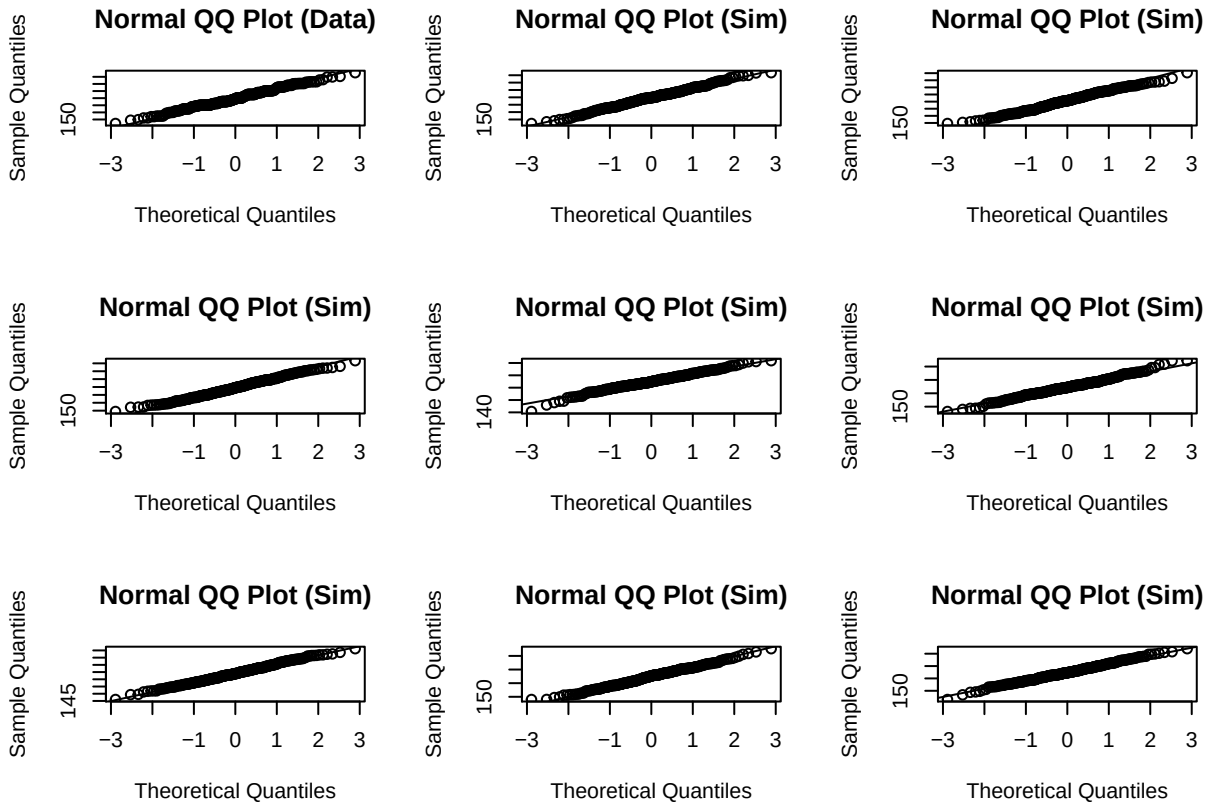
See below for the normal probability plot of 'sim_norm'. I also added the line to see how points fall relative to the normal distribution. The q-q plot for the simulted data is quite similar to the q-q plot for the real data - where many heights close to the mean and/or 1 standard deviation from the mean fall on or close to the line, and heights deviate a bit from the line as you get farther out at 2 to 3 standard deviations from the mean.

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



Even better than comparing the original plot to a single plot generated from a normal distribution is to compare it to many more plots using the following function. It may be helpful to click the zoom button in the plot window.

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



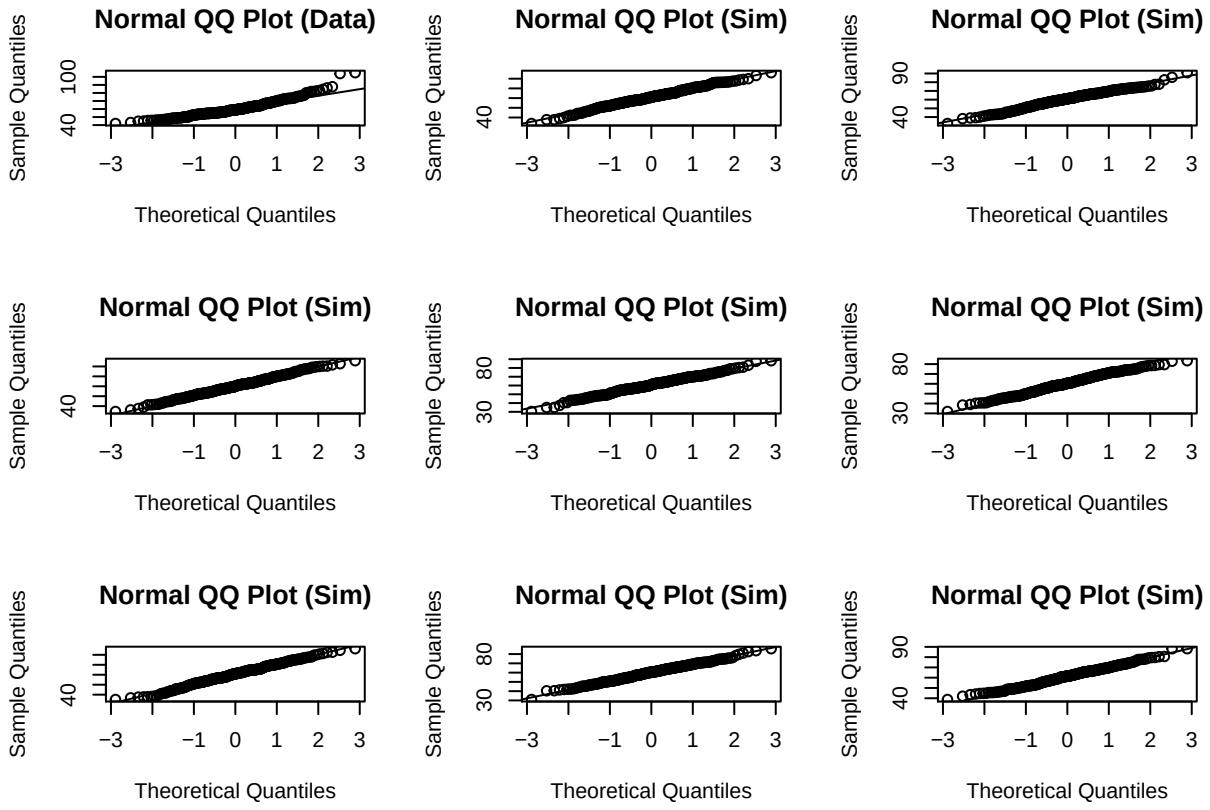
4. Does the normal probability plot for `fdims$hgt` look similar to the plots created for the simulated data? That is, do plots provide evidence that the female heights are nearly normal?

The probability plot for 'fdims$hgt' looks similar to the plots created for the simulated data, where heights fall on or near the line at the mean and on standard deviation from the mean (+/-), and then there is some deviation from the mean at 2 or 3 deviations from the mean (+/-), which are at the tails. The plots do provide evidence that the female heights are nearly normal.

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

From the comparison of the simulated data below with the female weights in the dataset, we can see that female weights do not follow as much of a normal distribution as female heights in this dataset. Although many weights are on or near the q-q plot line between the mean and one standard deviation from the mean, unlike the simulations where weights tend to still be close to the line between one and two standard deviations from the mean, the actual weight data tends to deviate from a normal distribution at this point, especially as you get past 1.5 deviations from the mean.

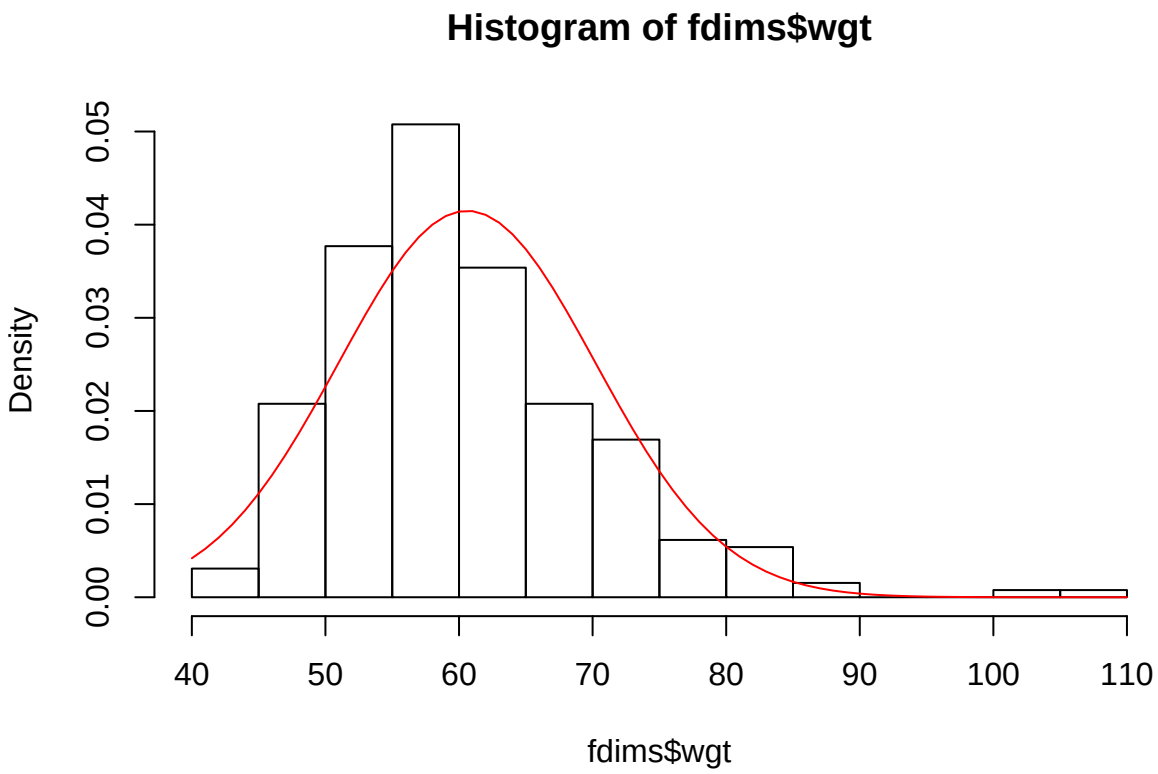
```
qqnormsim(fdims$wgt)
```



To get a better sense of the distribution, I also decided to plot it as a frequency distribution. See this below:

```
hist(fdims$wgt, probability = TRUE)
fwgtmean <- mean(fdims$wgt)
fwgtsd <- sd(fdims$wgt)

x <- 40:110
y <- dnorm(x = x, mean = fwgtmean, sd = fwgtsd)
lines(x = x, y = y, col = "red")
```



From the distribution plotted above, we can see that women's weights tend to be more right skewed. Visually, we can see that this deviates from a normal distribution.

Normal probabilities

Okay, so now you have a slew of tools to judge whether or not a variable is normally distributed. Why should we care?

It turns out that statisticians know a lot about the normal distribution. Once we decide that a random variable is approximately normal, we can answer all sorts of questions about that variable related to probability. Take, for example, the question of, "What is the probability that a randomly chosen young adult female is taller than 6 feet (about 182 cm)?" (The study that published this data set is clear to point out that the sample was not random and therefore inference to a general population is not suggested. We do so here only as an exercise.)

If we assume that female heights are normally distributed (a very close approximation is also okay), we can find this probability by calculating a Z score and consulting a Z table (also called a normal probability table). In R, this is done in one step with the function `pnorm`.

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

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

Note that the function `pnorm` gives the area under the normal curve below a given value, `q`, with a given mean and standard deviation. Since we're interested in the probability that someone is taller than 182 cm, we have to take one minus that probability.

Assuming a normal distribution has allowed us to calculate a theoretical probability. If we want to calculate the probability empirically, we simply need to determine how many observations fall above 182 then divide this number by the total sample size.

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

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

Although the probabilities are not exactly the same, they are reasonably close. The closer that your distribution is to being normal, the more accurate the theoretical probabilities will be.

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?

For my first set of probability questions regarding female heights, I was curious to see: What is the probability that a female in the dataset will have a height that is greater than 155 cm (~5ft 1in)?

Using the pnorm method:

```
1 - pnorm(q = 155, mean = fhgtmean, sd = fhgtsd)
```

```
## [1] 0.9342823
```

Calculating empirically:

```
sum(fdims$hgt > 155) / length(fdims$hgt)
```

```
## [1] 0.9307692
```

We can see that there is a strong likelihood (around a 93% chance) that a female will be taller than 155 centimeters in this dataset.

For my second set of probability questions regarding female weights, I was curious to see: What is the probability that a female in the dataset will have a weight that is greater than 90 kg (~198 lbs)?

Using the pnorm method:

```
1 - pnorm(q = 90, mean = fwgtmean, sd = fwgtsd)
```

```
## [1] 0.001116107
```

Calculating empirically:

```
sum(fdims$wgt > 90) / length(fdims$wgt)
```

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

We can see that there is a small likelihood (less than a 1% chance) that a female will weigh more than 90 kg in this dataset.

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.

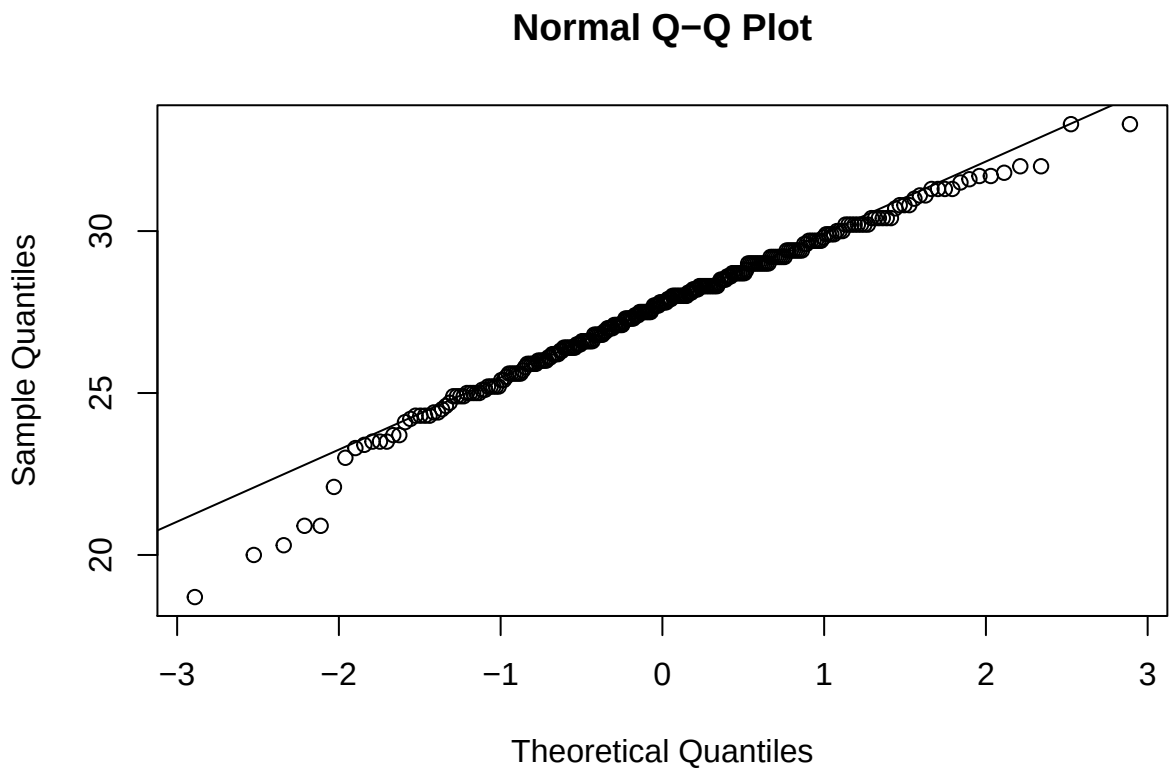
  a. The histogram for female biiliac (pelvic) diameter (`bii.di`) belongs to normal probability plot letter **B**.

  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**.

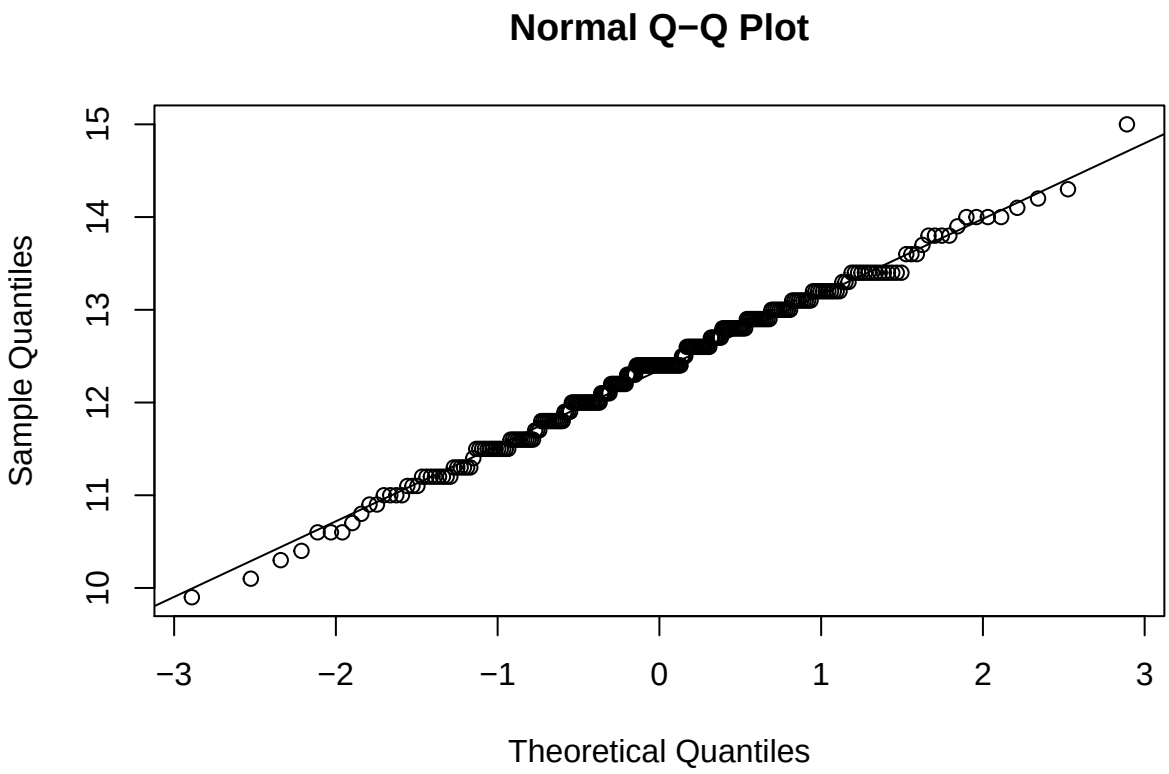
  d. The histogram for female chest depth (`che.de`) belongs to normal probability plot letter **A**.

```
# check a.
qqnorm(fdims$bii.di)
qqline(fdims$bii.di)
```



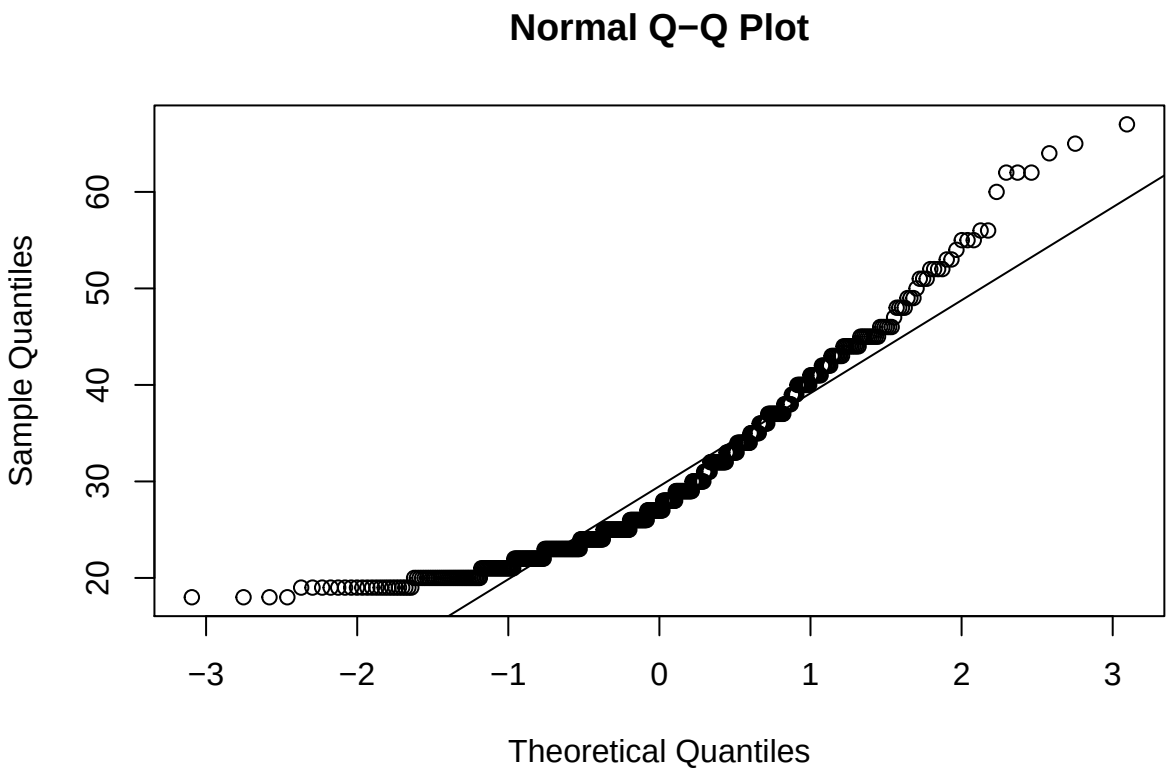

```
# confirmed that this matches with plot B.
```

```
# check b.
qqnorm(fdims$elb.di)
qqline(fdims$elb.di)
```



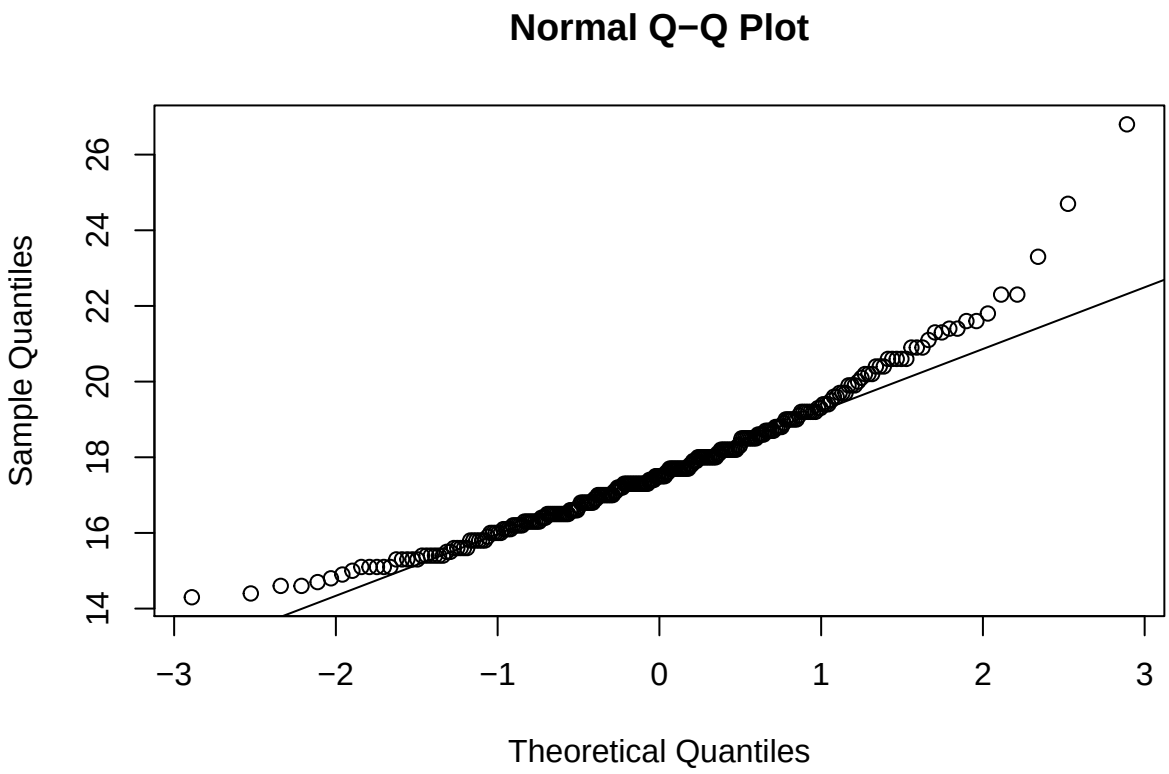

```
# confirmed that this matches with plot C.
```

```
# check c.
qqnorm(bdims$age)
qqline(bdims$age)
```




```
# confirmed that this matches with plot D.
```

```
# check d.
qqnorm(fdims$che.de)
qqline(fdims$che.de)
```




```
# confirmed that this matches with plot A.
```

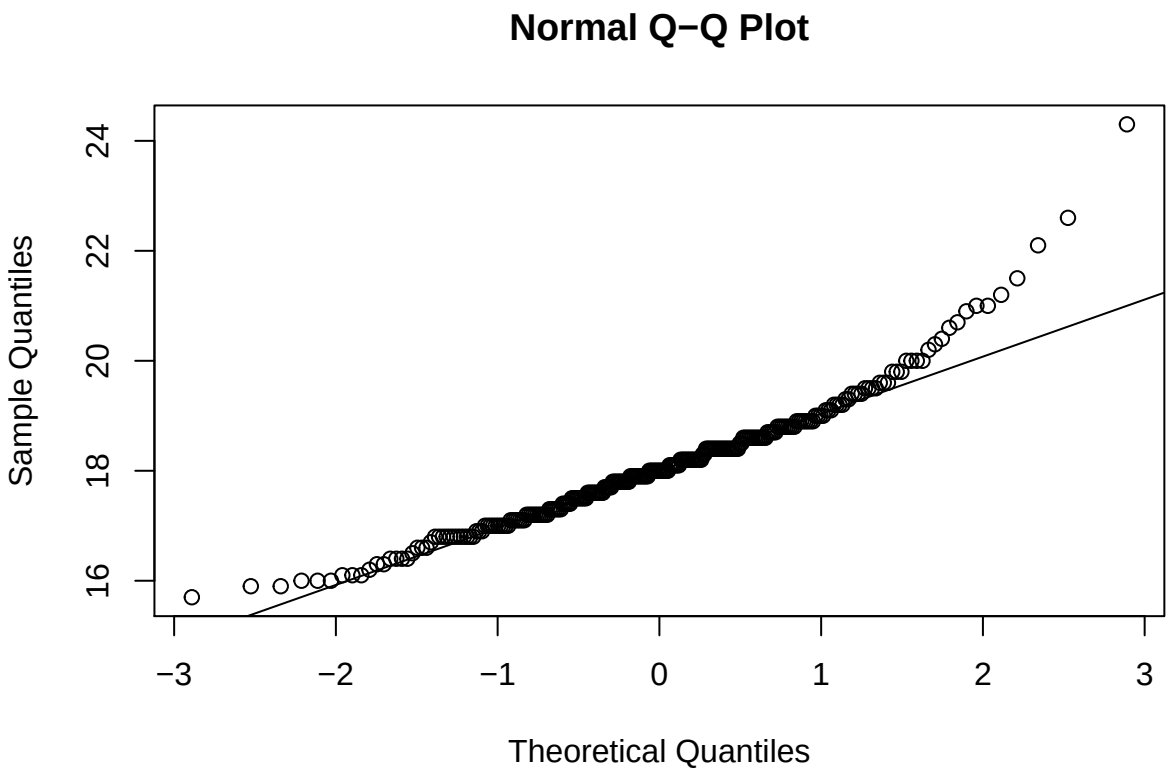
- Note that normal probability plots C and D have a slight stepwise pattern.
  Why do you think this is the case?

The reason why plots C and D have a slight stepwise pattern is due to data values for age and female elbow diameter. The age data contains ages rounded to the nearest whole number. Therefore, when viewing this data on a continuous scale, there will not be any decimal values, making it display as stepwise. For female elbow diameter, the paper/methodology states that this value is the sum of the two elbows. Therefore, there may have been some rounding that occurred during this calculation, which would have resulted in this stepwise pattern.

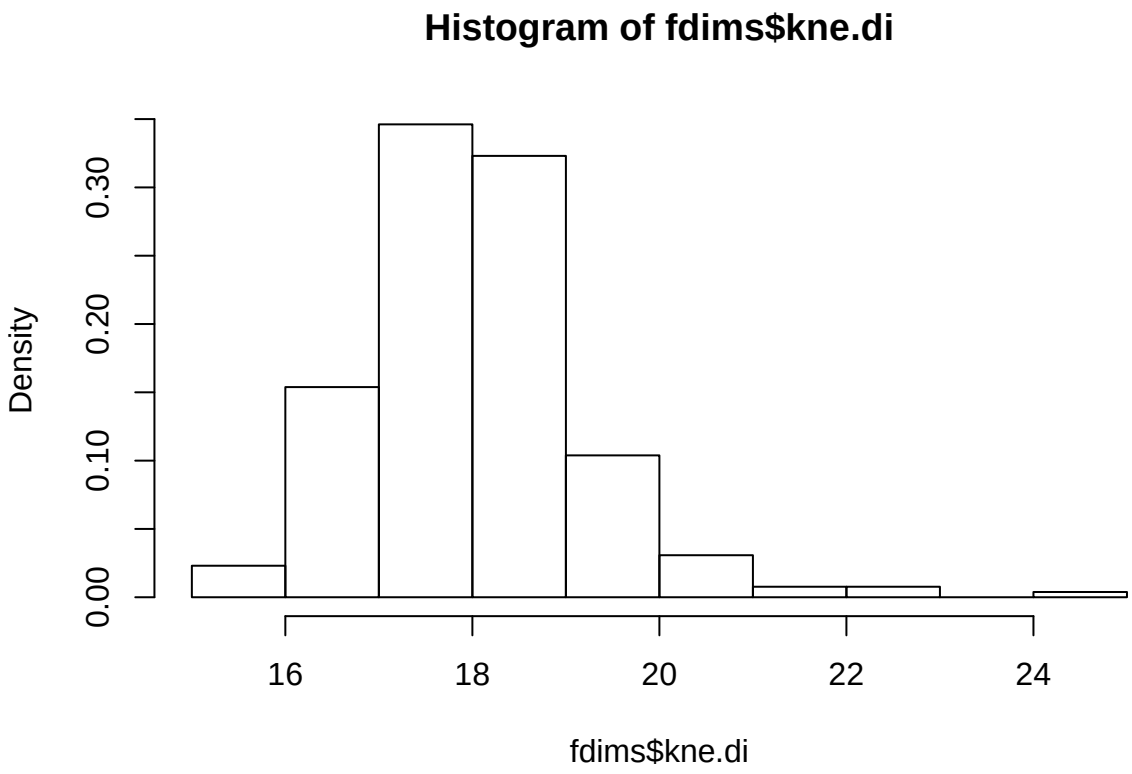
- As you can see, normal probability plots can be used both to assess normality and visualize skewness. Make a normal probability plot for female knee diameter (`kne.di`). Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Use a histogram to confirm your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

From the normal probability plot, this variable of female knee diameter appears to be right skewed, given that many values deviate from the qqline from 1.5 standard deviations from the mean and beyond.

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




```
hist(fdims$kne.di, probability = TRUE)
```




```
fknemean <- mean(fdims$kne.di)
fknestd <- sd(fdims$kne.di)
```

This is confirmed by the histogram, as we can see that the distribution is right skewed with a right tail.

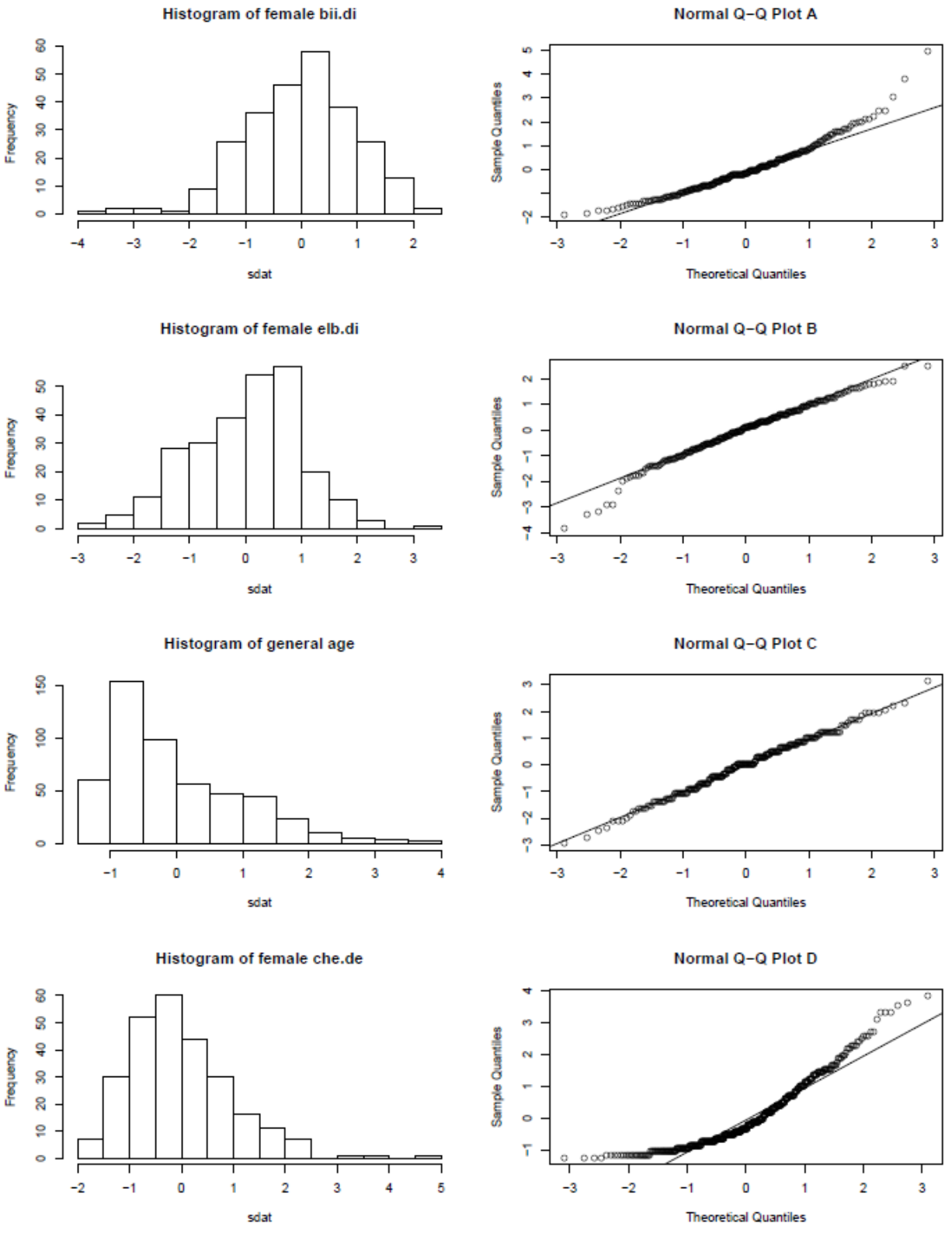


Figure 1: histQQmatch