Lab3 - The normal distribution

Michael Y.

October 7th, 2018

```
setwd("c:/users/Michael/DROPBOX/priv/CUNY/MSDS/201809-Fall/DATA606_Jason/Labs/Lab3")
```

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
```

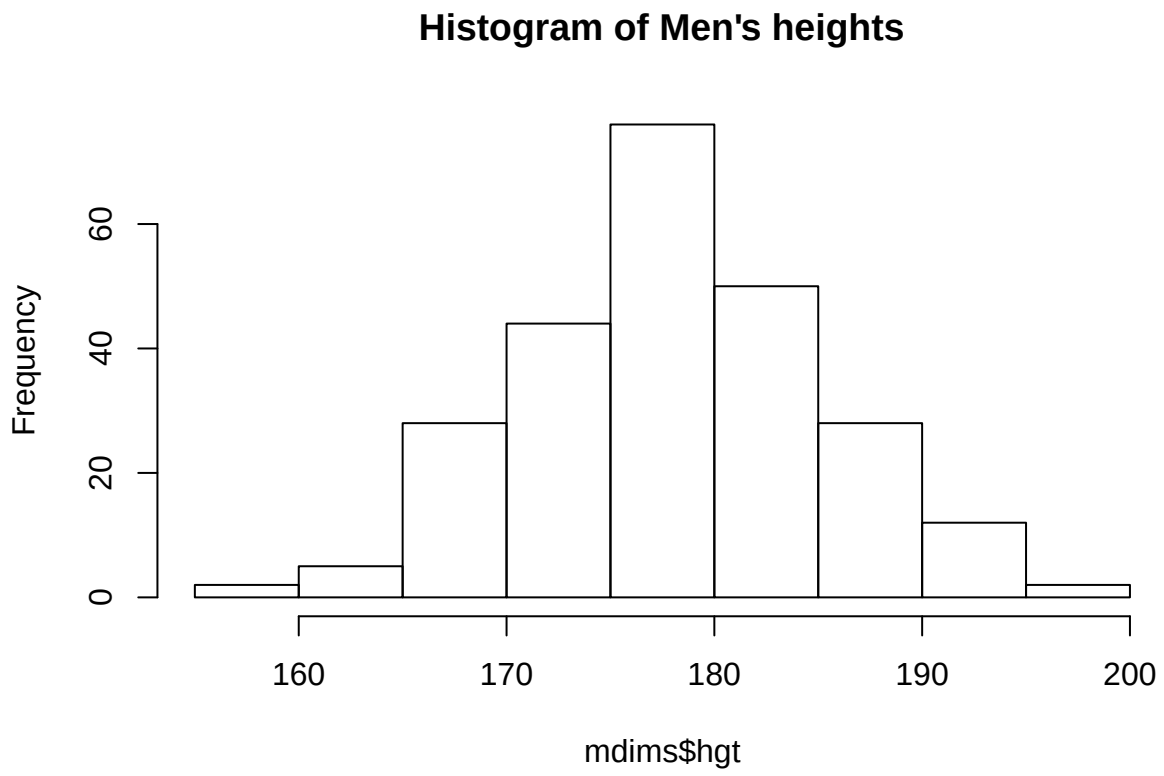
You'll see that for every observation we have 25 measurements, many of which are either diameters or girths. A key to the variable names can be found at http://www.openintro.org/stat/data/bdims.php, but we'll be focusing on just three columns to get started: weight in kg (`wgt`), height in cm (`hgt`),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

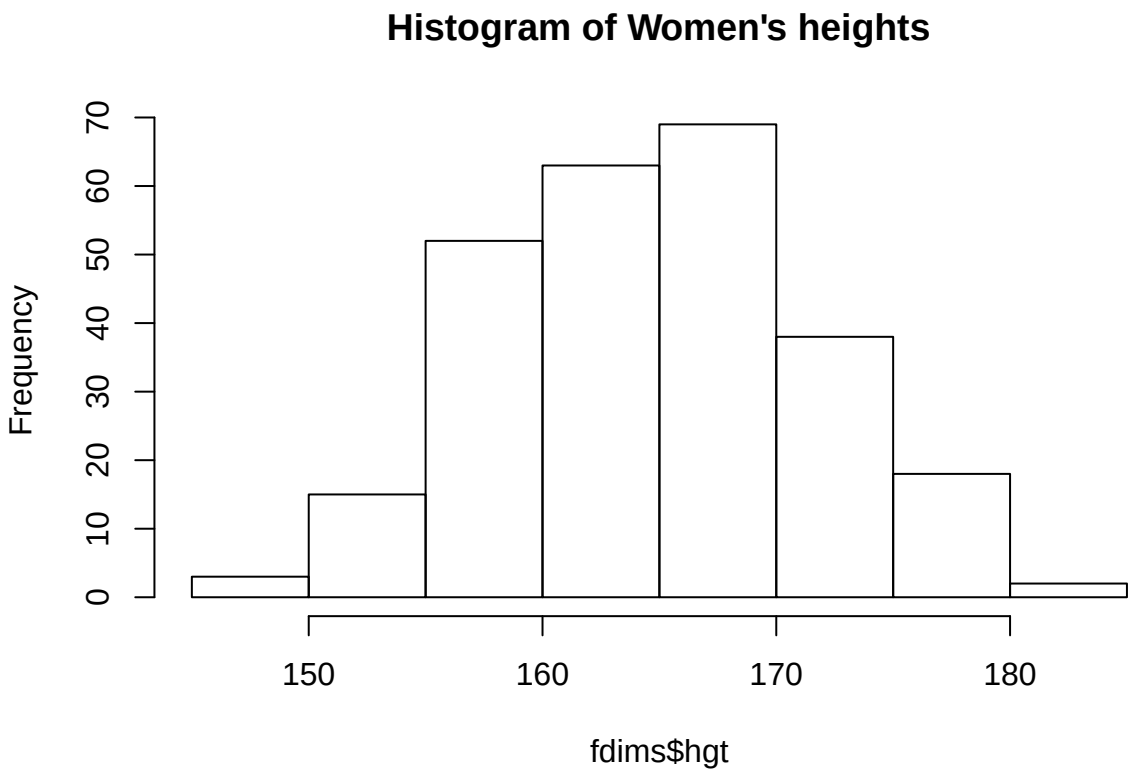
```r
mdims <- subset(bdims, sex == 1)
fdims <- subset(bdims, sex == 0)
```

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?

```r
hist(mdims$hgt,main = "Histogram of Men's heights")
```




```r
hist(fdims$hgt,main = "Histogram of Women's heights")
```



The men's histogram reflects a higher median than the women's (178cm vs. 165cm).

Both distributions are unimodal and close to symmetric – they appear to be roughly Normal.

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.

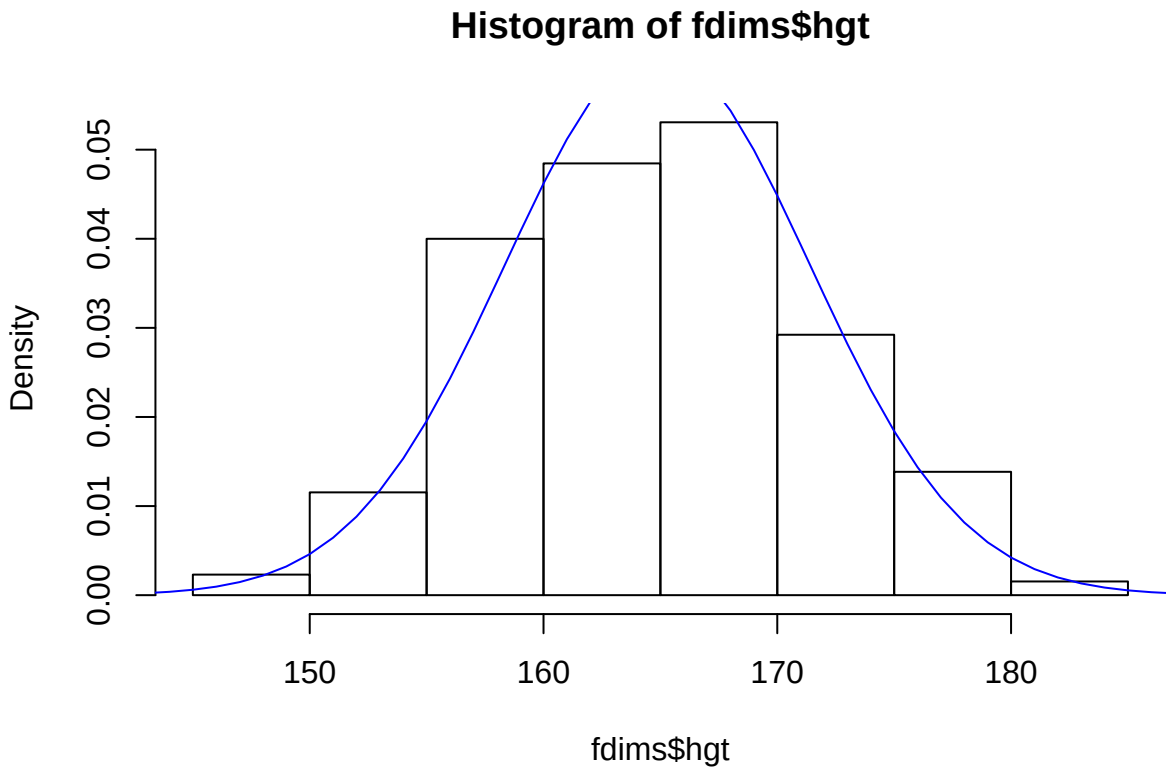
To see how accurate that description is, we can plot a normal distribution curve on top of a histogram to see how closely the data follow a normal distribution. This normal curve should have the same mean and standard deviation as the data. We'll be working with women's heights, so let's store them as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

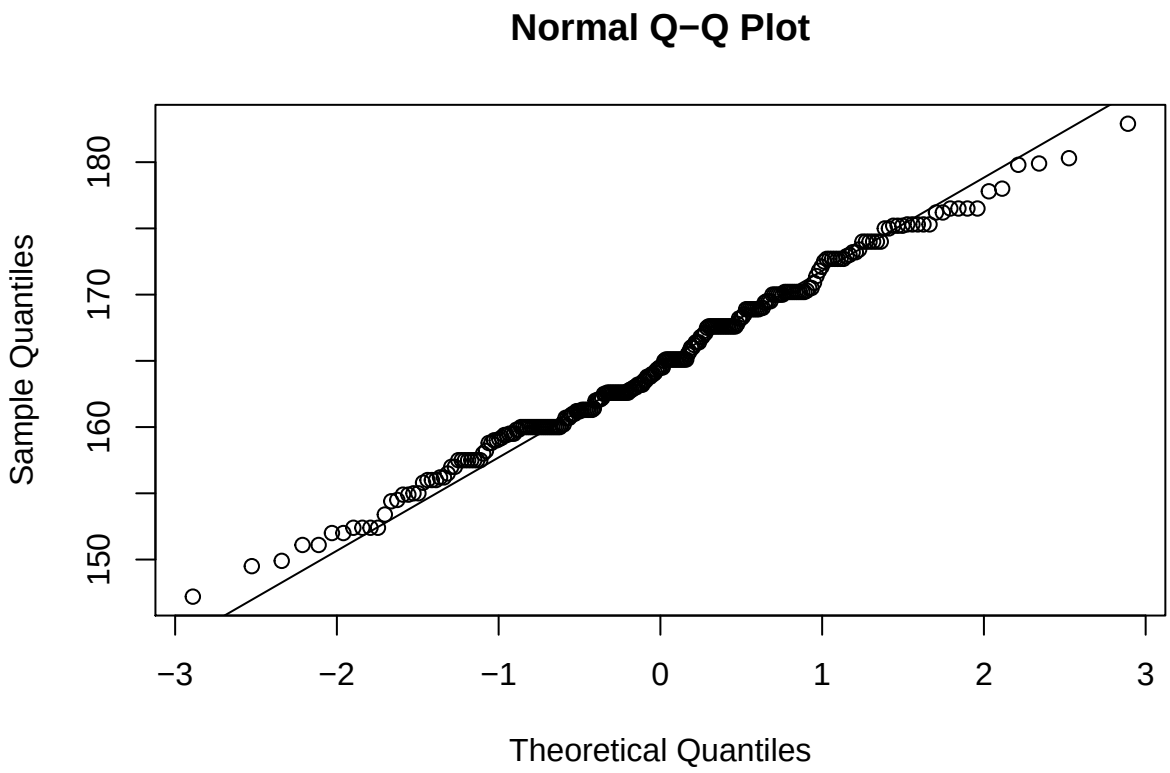
2. Based on the this plot, does it appear that the data follow a nearly normal distribution?

Yes.

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?

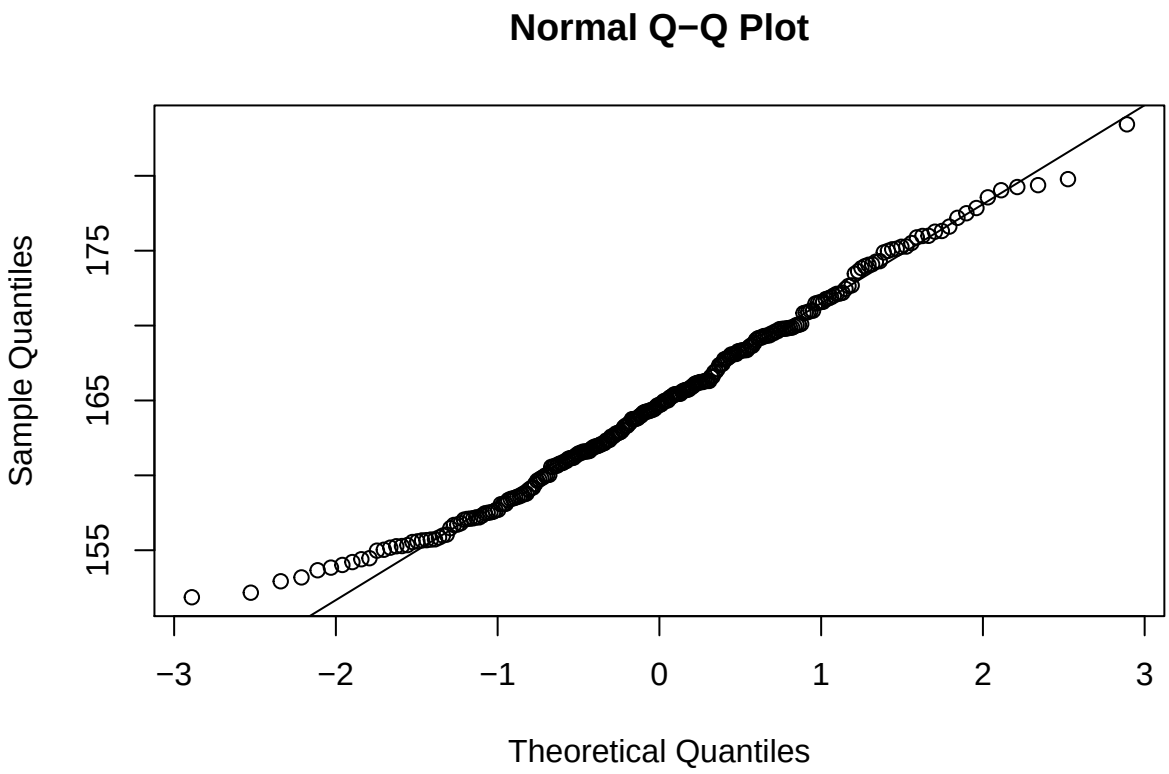
A useful way to address this question is to rephrase it as: what do probability plots look like for data that I *know* came from a normal distribution? We can answer this by simulating data from a normal distribution using `rnorm`.

```
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?

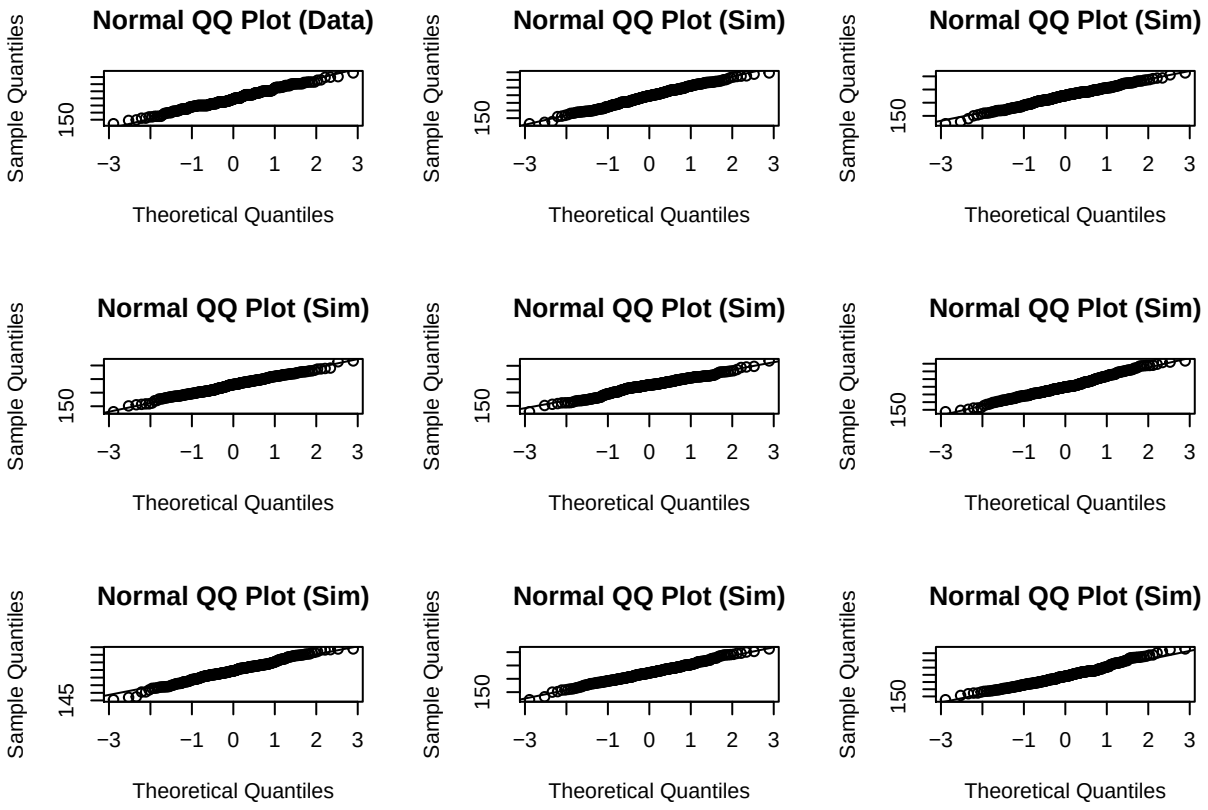
```
qqnorm(sim_norm)
qqline(sim_norm)
```



The points near the center of the QQ-plot tend to fall on the line, but the extreme points tend to deviate slightly from the line. Under the simulation that generated the above plot, the points on the right were above the line while the points on the left fell below the line. (Note that as this Rmarkdown file is repeatedly re-evaluated/knit, each simulation may differ from what I have written here.) With regard to the actual data, the points on the right fall below the line while the points on the left are above the line.

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?

Yes, the above simulated plots reflect 9 repetitions of the simulation, where the extreme values fluctuate from falling slightly above vs. slightly below the line. These are, overall, consistent with the actual female heights.

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

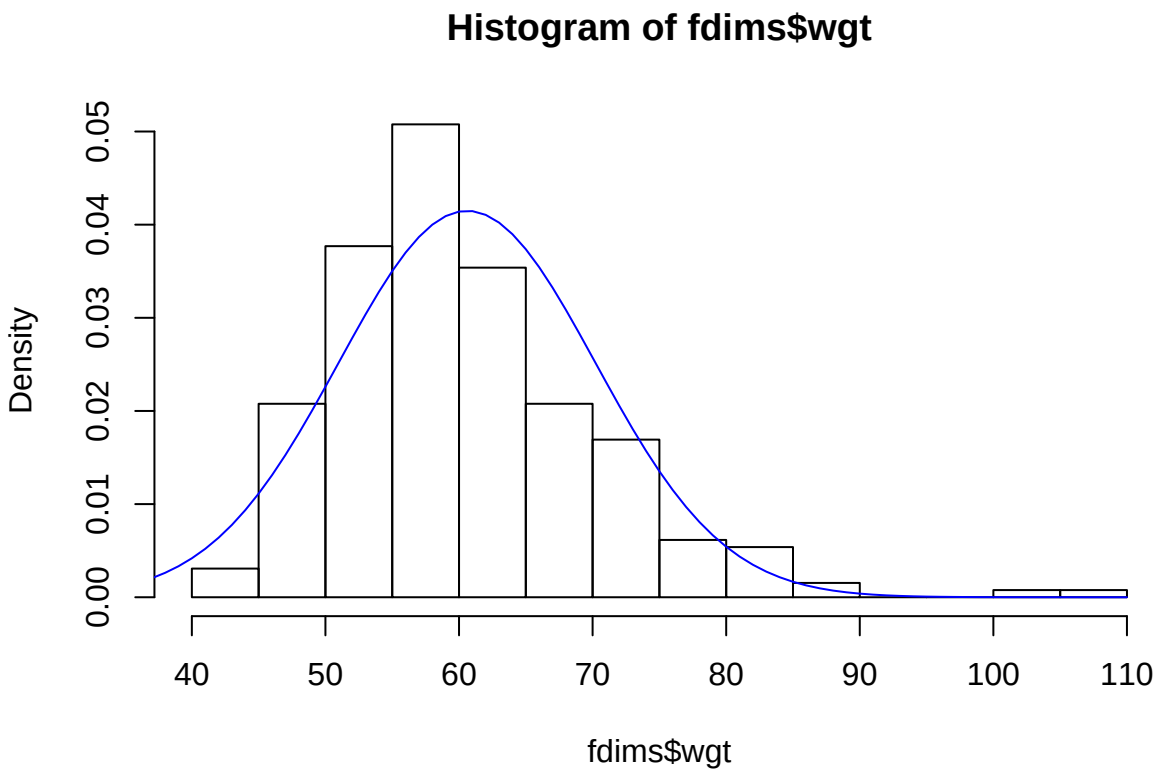
```
fwgtmean <- mean(fdims$wgt)
fwgtsd   <- sd(fdims$wgt)
fwgtsd
```

```
## [1] 9.615699
```

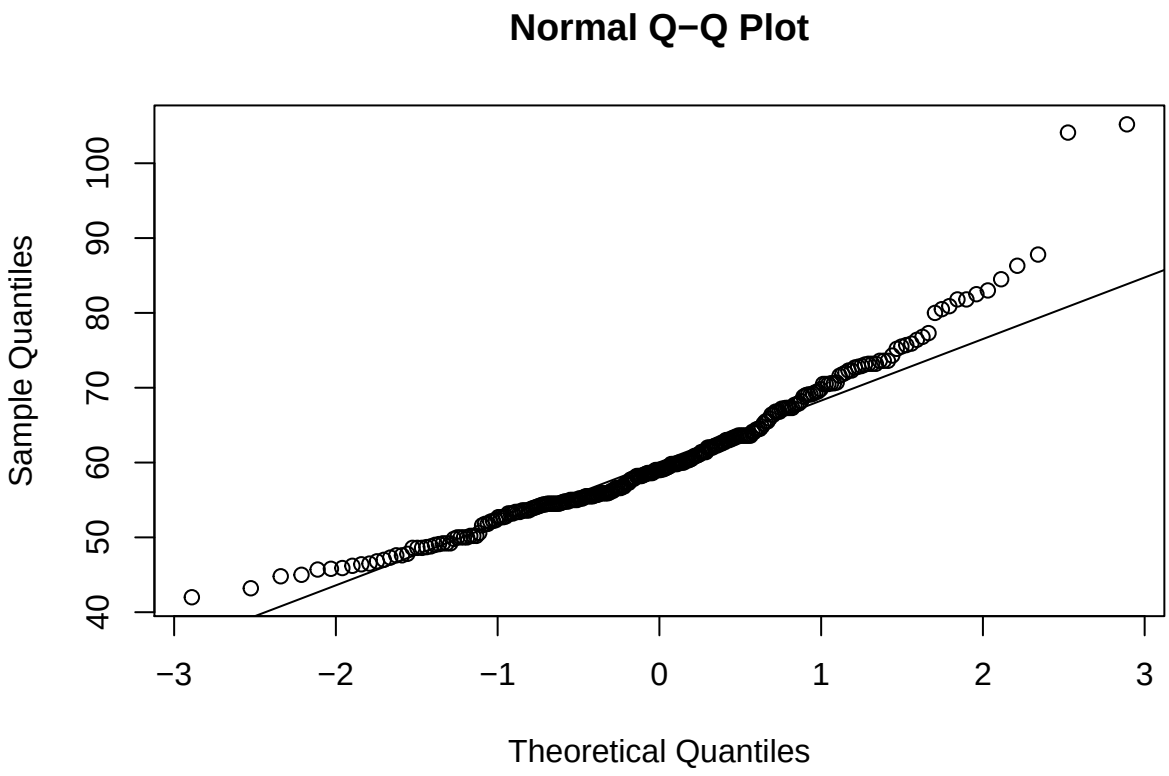
```
summary(fdims$wgt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    42.0    54.5    59.0    60.6    65.6   105.2
```

```
hist(fdims$wgt, probability = TRUE)
x <- 30:110
y <- dnorm(x = x, mean = fwgtmean, sd = fwgtsd)
lines(x = x, y = y, col = "blue")
```




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



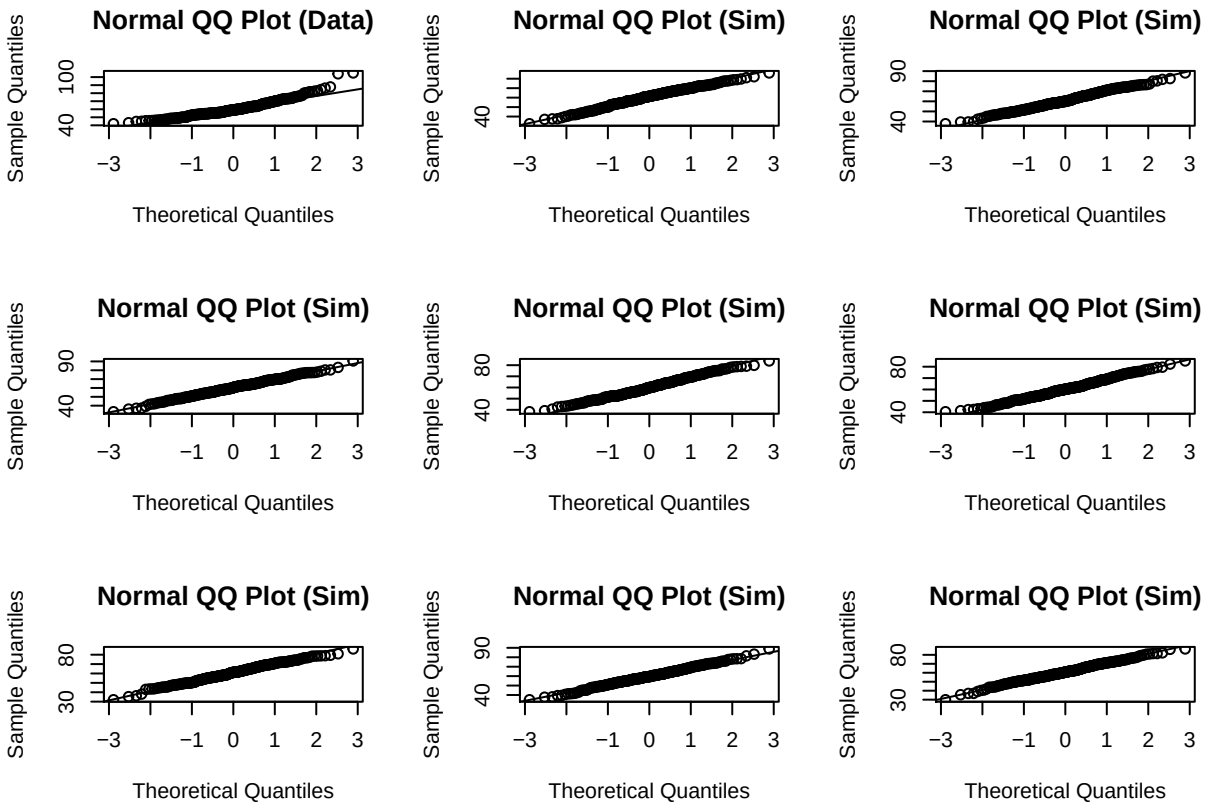

```
sim_norm_fwgt <- rnorm(n = length(fdims$wgt), mean = fwgtmean, sd = fwgtsd)
sd(sim_norm_fwgt)
```

```
## [1] 9.021506
```

```
summary(sim_norm_fwgt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   39.34   56.47   61.73   62.38   69.13   85.06
```

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



In contrast to the heights, the actual distribution of weights appears to be skewed to the right – i.e., there are a number of outliers creating a "fat tail" (no pun intended?). The simulated data does not exhibit this pattern. For a true Normal distribution, the mean and the median should be equal. In the actual weights data, the mean is 1.6kg greater than the median (60.6 vs. 59) which is further evidence of a right skew.

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?

(a) What is the probability that a randomly chosen young adult female is less than 150cm tall?

Theoretial distribution:

```
pnorm(q = 150, mean = fhgtmean, sd = fhgtsd)
```

```
## [1] 0.01152955
```

This is approximately 3 out of 260.

Empirical distribution:

```
sum(fdims$hgt < 150) / length(fdims$hgt)
```

```
## [1] 0.01153846
```

This is exactly 3 out of 260.

For height, the empirical distribution agrees extremely closely with the theoretical distribution for short women.

Weight

(b) What is the probability that a randomly chosen young adult female weighs more than 87kg?

Theoretial distribution:

```
result = 1-pnorm(q = 87, mean = fwgtmean, sd = fwgtsd)
result
```

```
## [1] 0.003021211
```

```
result*260
```

```
## [1] 0.785515
```

This is less than 1 out of 260.

Empirical distribution:

```
sum(fdims$wgt > 87) / length(fdims$wgt)
```

```
## [1] 0.01153846
```

This is exactly 3 out of 260.

For weight, the empirical distribution contains a significantly larger number of heavy women than does the theoretical distribution.

Result: The distrivution of heights exhibits much closer agreement between the empirical and theoretical methods.

On Your Own

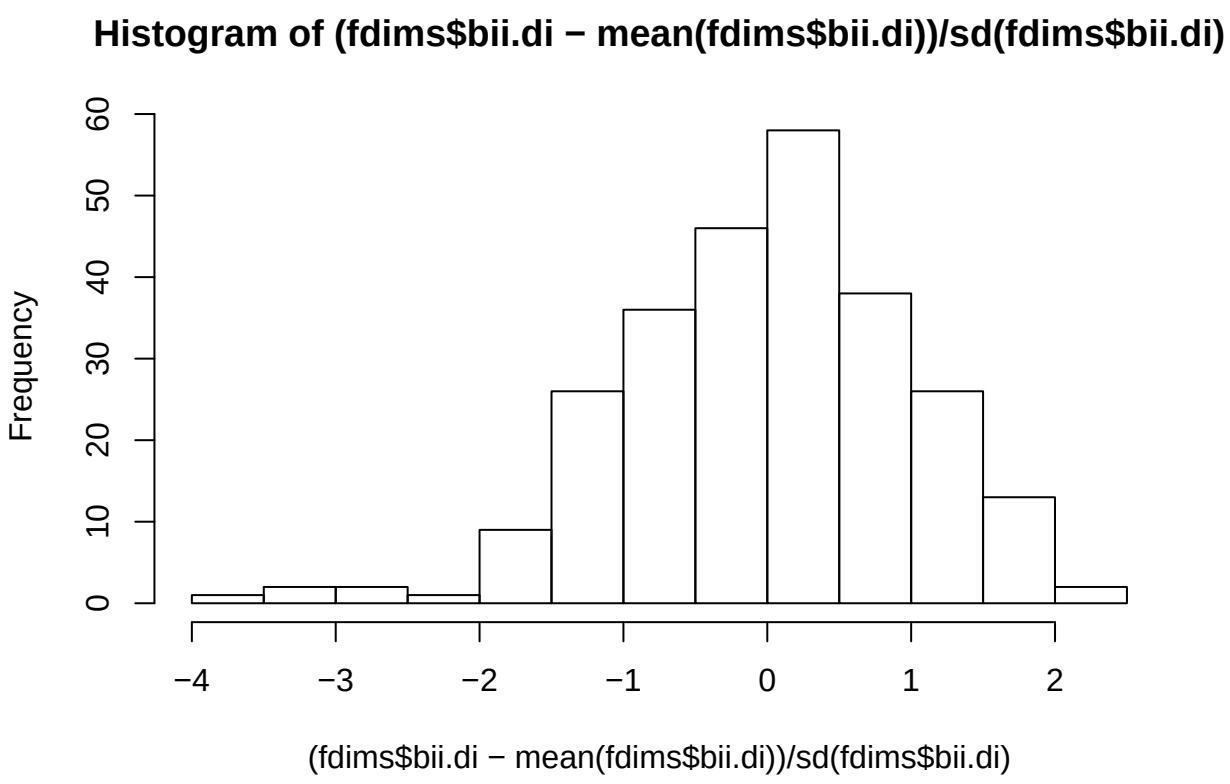
(1) Now let's consider some of the other variables in the body dimensions data set.

Using the figures at the end of the exercises, match the histogram to its normal probability plot.

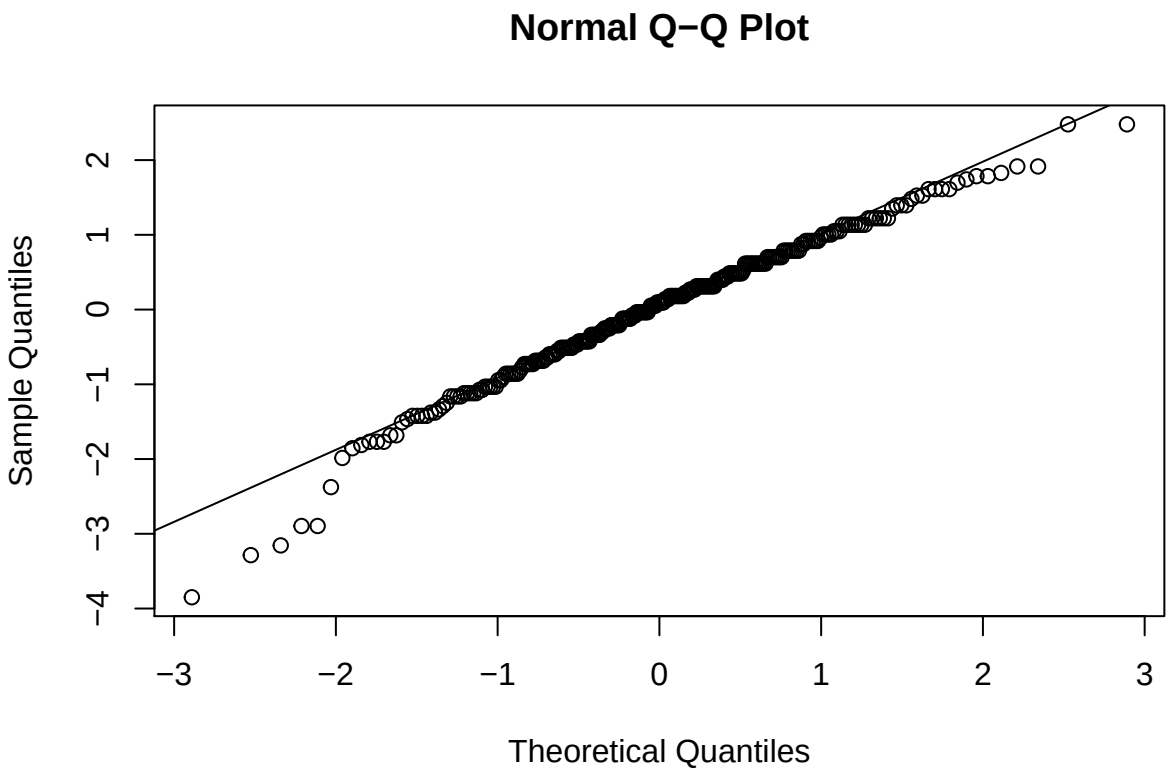
All of the variables have been standardized (first subtract the mean, then divide by the standard deviation), so the units won't be of any help.

If you are uncertain based on these figures, generate the plots in R to check.

```
hist((fdims$bii.di-mean(fdims$bii.di))/sd(fdims$bii.di),breaks = 13)
```



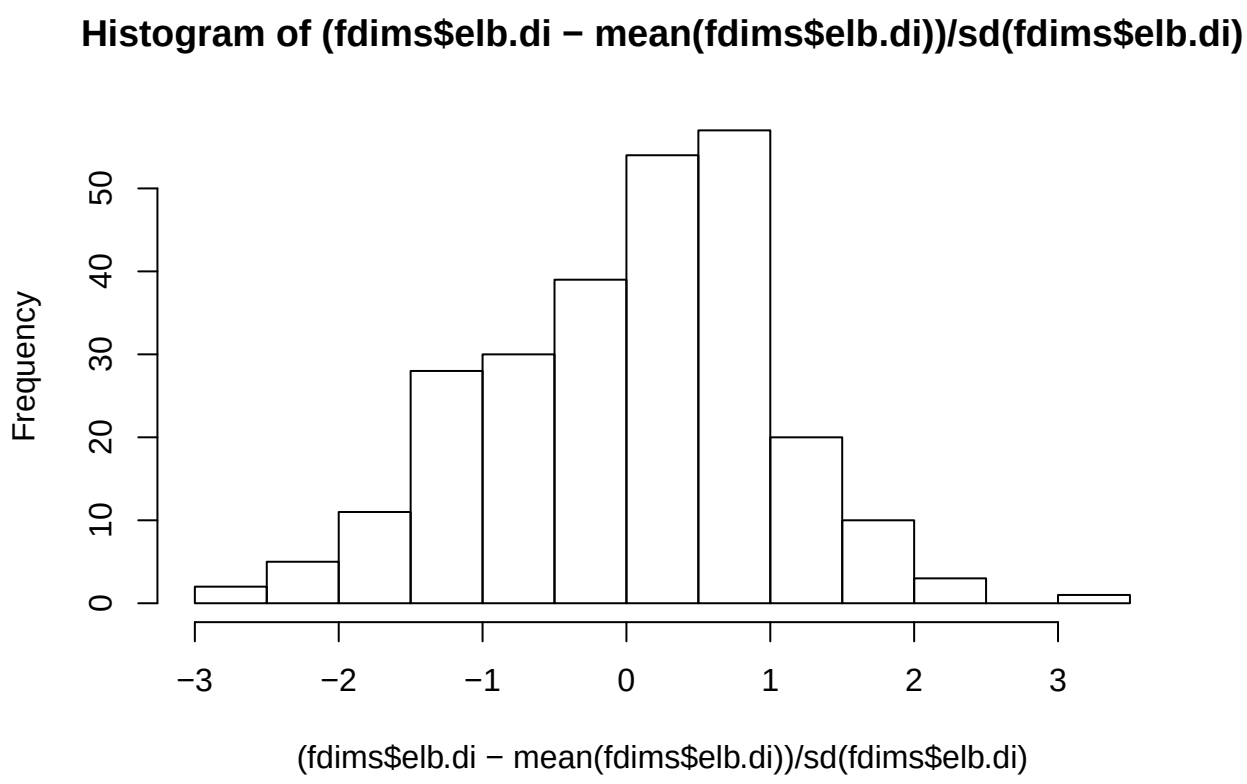

```
qqnorm((fdims$bii.di-mean(fdims$bii.di))/sd(fdims$bii.di))
qqline((fdims$bii.di-mean(fdims$bii.di))/sd(fdims$bii.di))
```



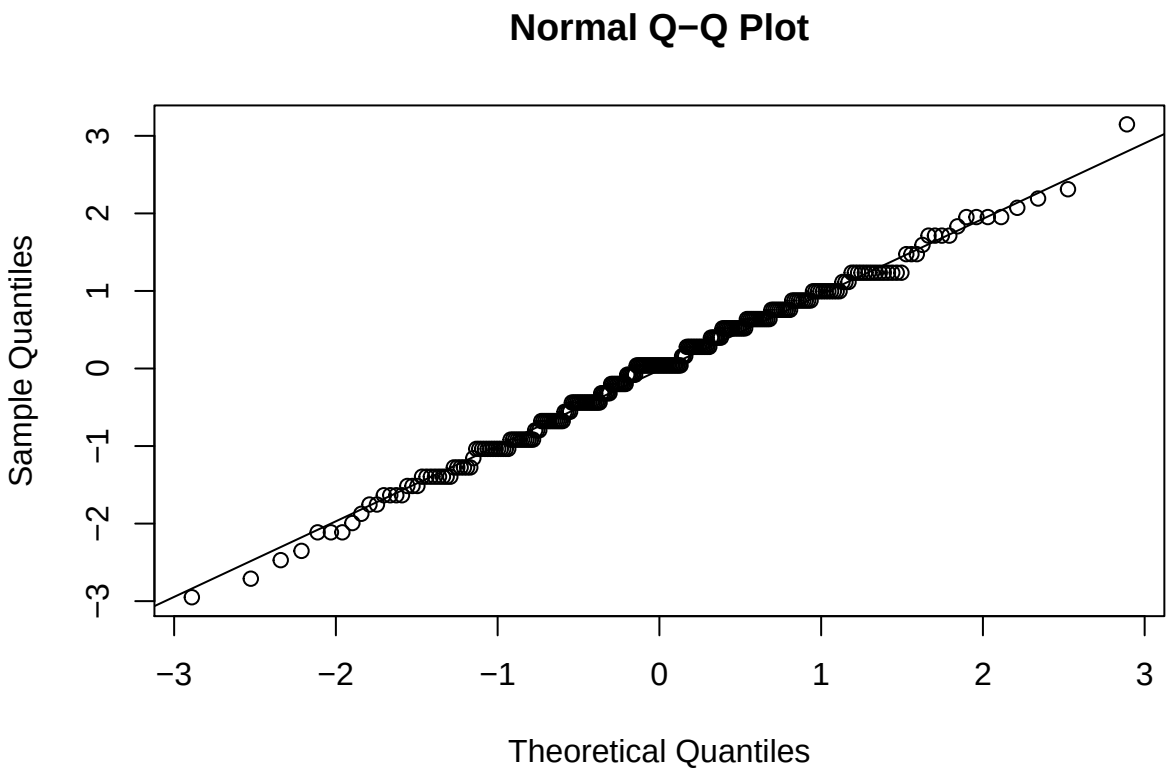
a. The histogram for female biiliac (pelvic) diameter (`bii.di`) belongs to normal probability plot letter B.

* *

```
hist((fdims$elb.di-mean(fdims$elb.di))/sd(fdims$elb.di),breaks = 13)
```



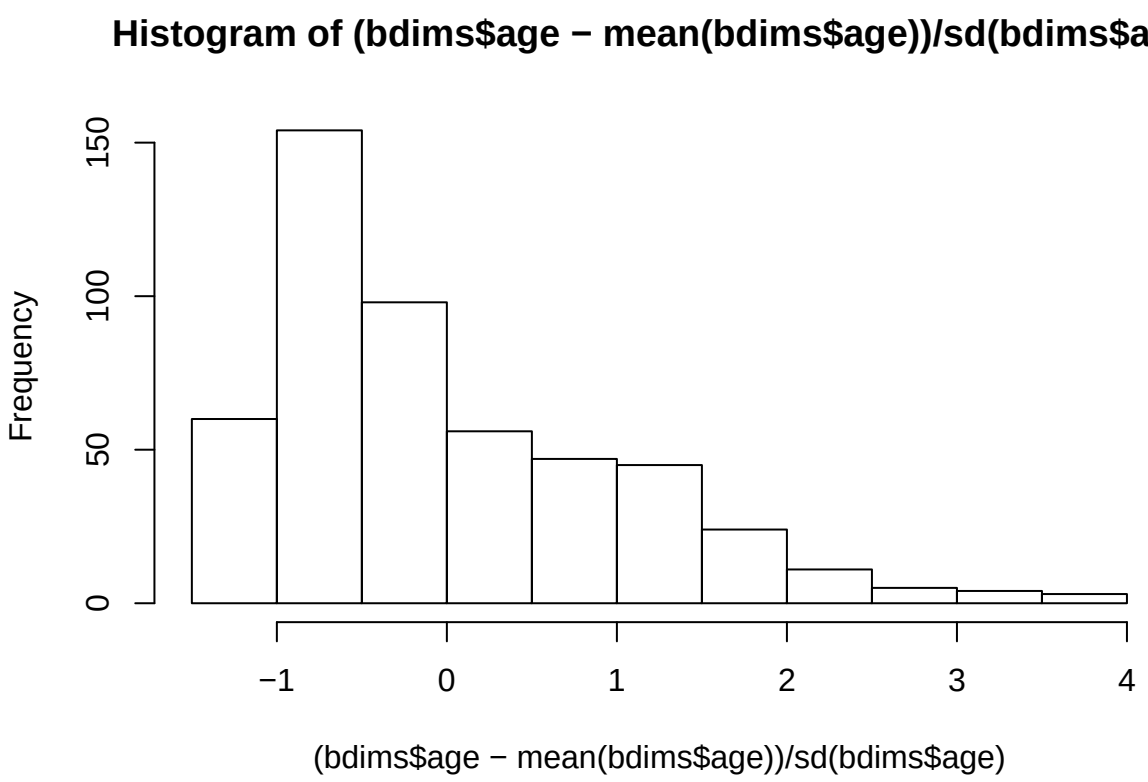

```
qqnorm((fdims$elb.di-mean(fdims$elb.di))/sd(fdims$elb.di))
qqline((fdims$elb.di-mean(fdims$elb.di))/sd(fdims$elb.di))
```



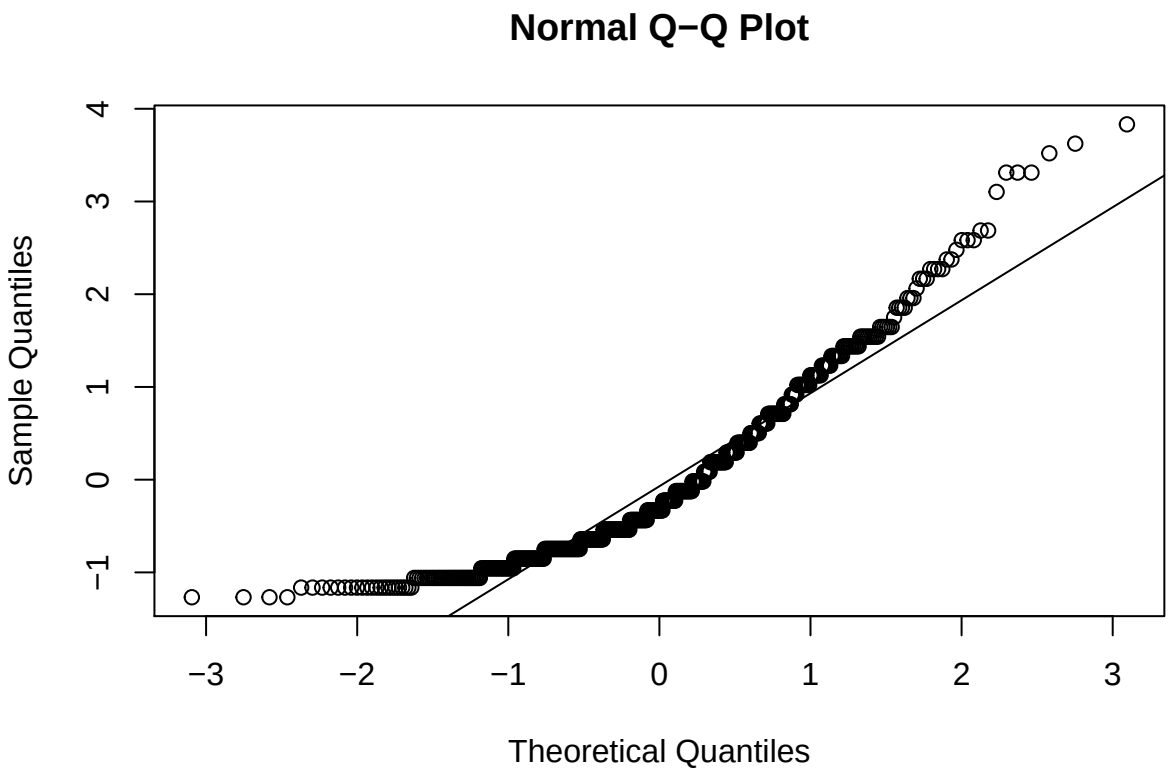
b. The histogram for female elbow diameter (`elb.di`) belongs to normal probability plot letter C.

* *

```
hist((bdims$age-mean(bdims$age))/sd(bdims$age),breaks = 13)
```



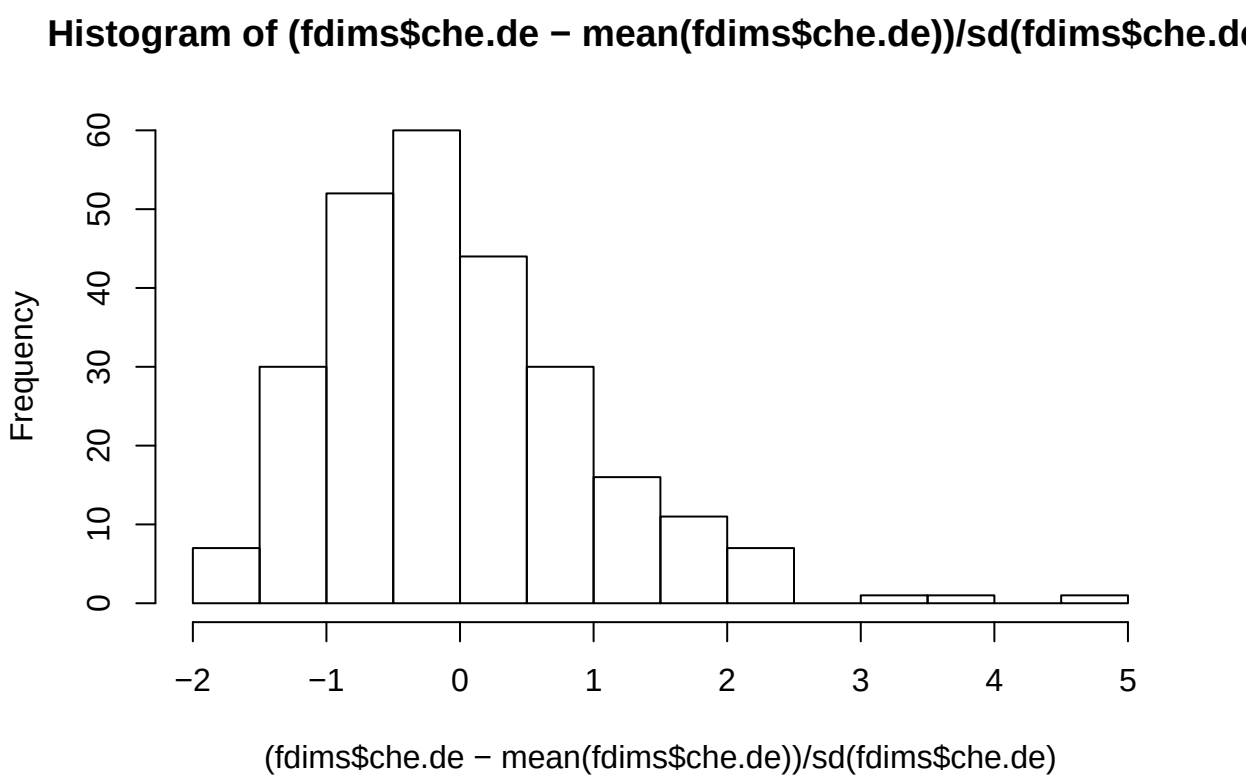

```
qqnorm((bdims$age-mean(bdims$age))/sd(bdims$age))
qqline((bdims$age-mean(bdims$age))/sd(bdims$age))
```



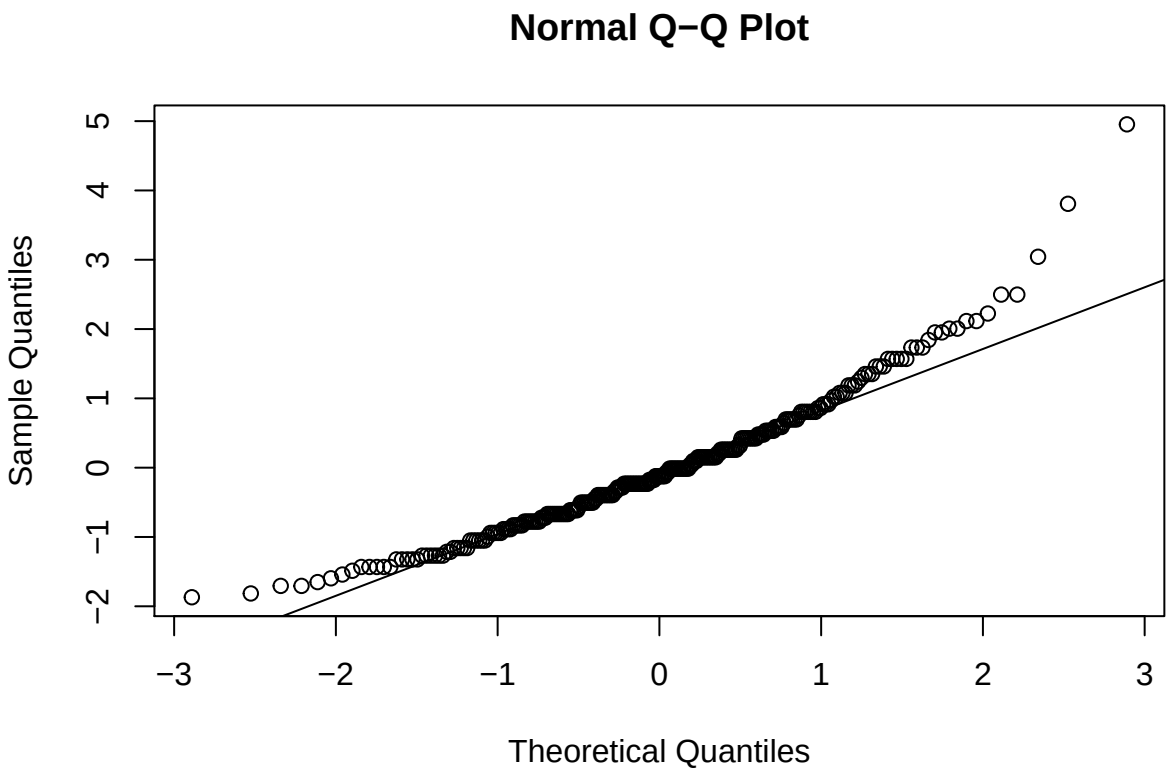
c. The histogram for general age (age) belongs to normal probability plot letter D.

* *

```
hist((fdims$che.de-mean(fdims$che.de))/sd(fdims$che.de),breaks = 13)
```




```
qqnorm((fdims$che.de-mean(fdims$che.de))/sd(fdims$che.de))
qqline((fdims$che.de-mean(fdims$che.de))/sd(fdims$che.de))
```



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

*** ***

(2) Note that normal probability plots C and D have a slight stepwise pattern.

Why do you think this is the case?

The stepwise pattern is seen when discrete values are measured many times. For example, plot C measures age, which in this data set is given in full years. Here there are only 44 distinct values for age, out of 507 entries. Many people at the same age will cause this stepwise behavior.

```
## table of ages
table(bdims$age)
```

```
##
## 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
##  4 22 34 25 28 39 28 34 23 21 17 22 15  8 20 13 15 10  9 16  7  5 11  9  7
## 43 44 45 46 47 48 49 50 51 52 53 54 55 56 60 62 64 65 67
##  8 10 10  6  1  4  3  1  3  4  2  1  3  2  1  3  1  1  1
```

For chest depth, there are a larger number of measurements used (67) but certain measurements are used much more frequently. This could be a limitation of the measuring device. Again, stepwise behavior is observed.

```
###table of chest depth
table(fdims$che.de)
```

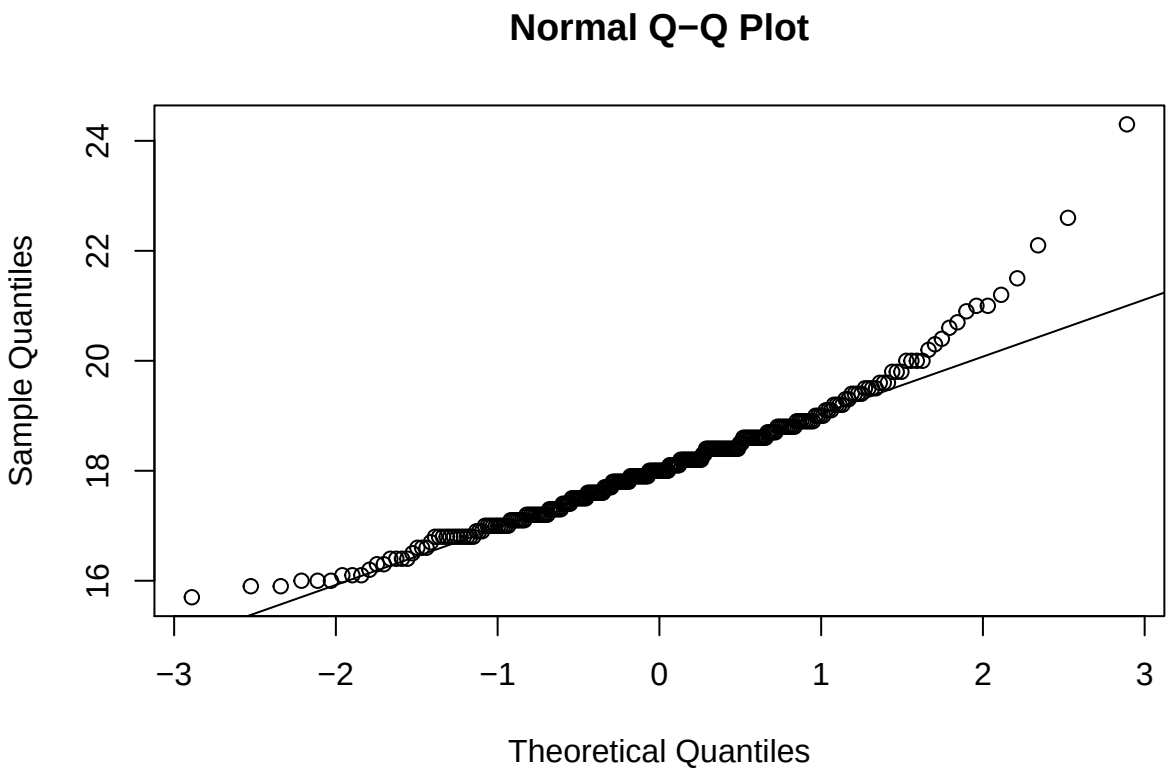
```
##
## 14.3 14.4 14.6 14.7 14.8 14.9   15 15.1 15.3 15.4 15.5 15.6 15.8 15.9   16
##    1    1    2    1    1    1    1    5    5    6    2    5    6    1    5
## 16.1 16.2 16.3 16.4 16.5 16.6 16.7 16.8 16.9   17 17.1 17.2 17.3 17.4 17.5
##    4    5    8    3   12    5    1    8    2   10    2    4   16    4    7
## 17.6 17.7 17.8 17.9   18 18.1 18.2 18.3 18.5 18.6 18.7 18.8 18.9   19 19.1
##    2   13    2    3   12    2   10    2    9    4    5    4    1    6    1
## 19.2 19.3 19.4 19.5 19.6 19.7 19.9   20 20.1 20.2 20.4 20.6 20.9 21.1 21.3
##    7    2    3    1    2    3    3    1    1    3    3    5    3    1    2
## 21.4 21.6 21.8 22.3 23.3 24.7 26.8
##    2    2    1    2    1    1    1
```

* *

(3) 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`).

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



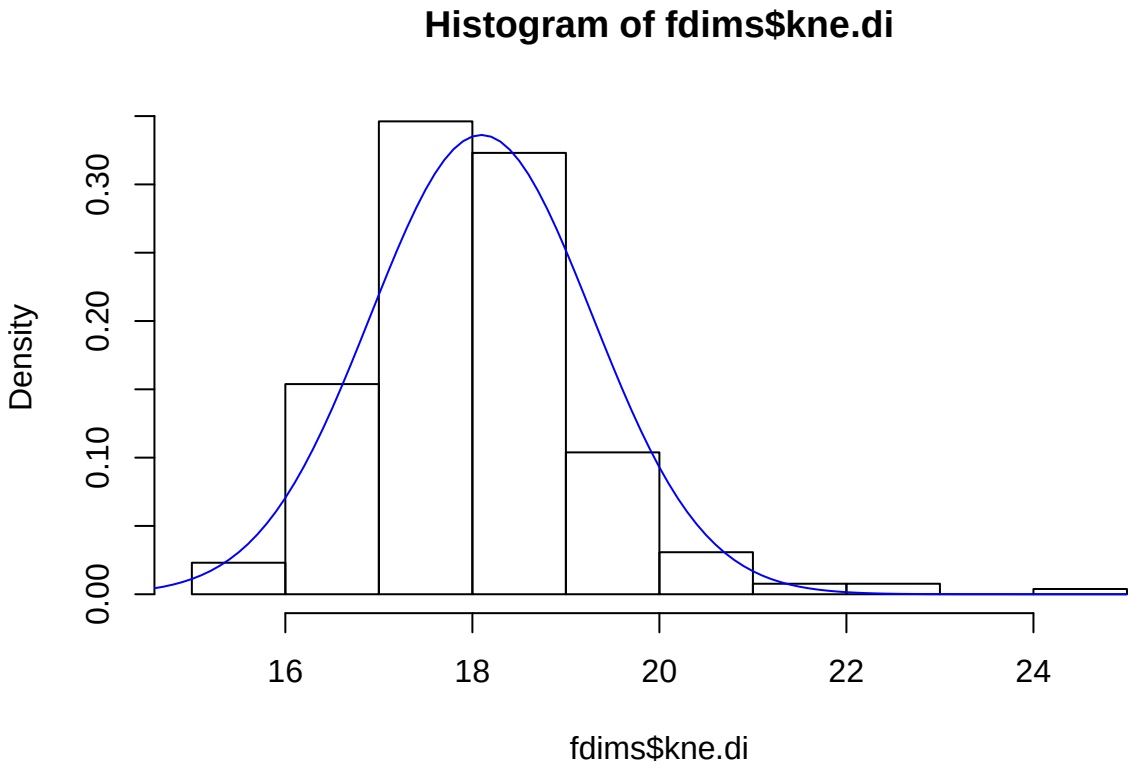
Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed?

The substantial number of outliers above the line on the far right indicate that this distribution is right-skewed.

Use a histogram to confirm your findings.

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

hist(fdims$kne.di, probability = TRUE)
x <- seq(from=12,to=25,by=0.1)
y <- dnorm(x = x, mean = mean(fdims$kne.di), sd = sd(fdims$kne.di))
lines(x = x, y = y, col = "blue")
```



Right-skew confirmed.

* *

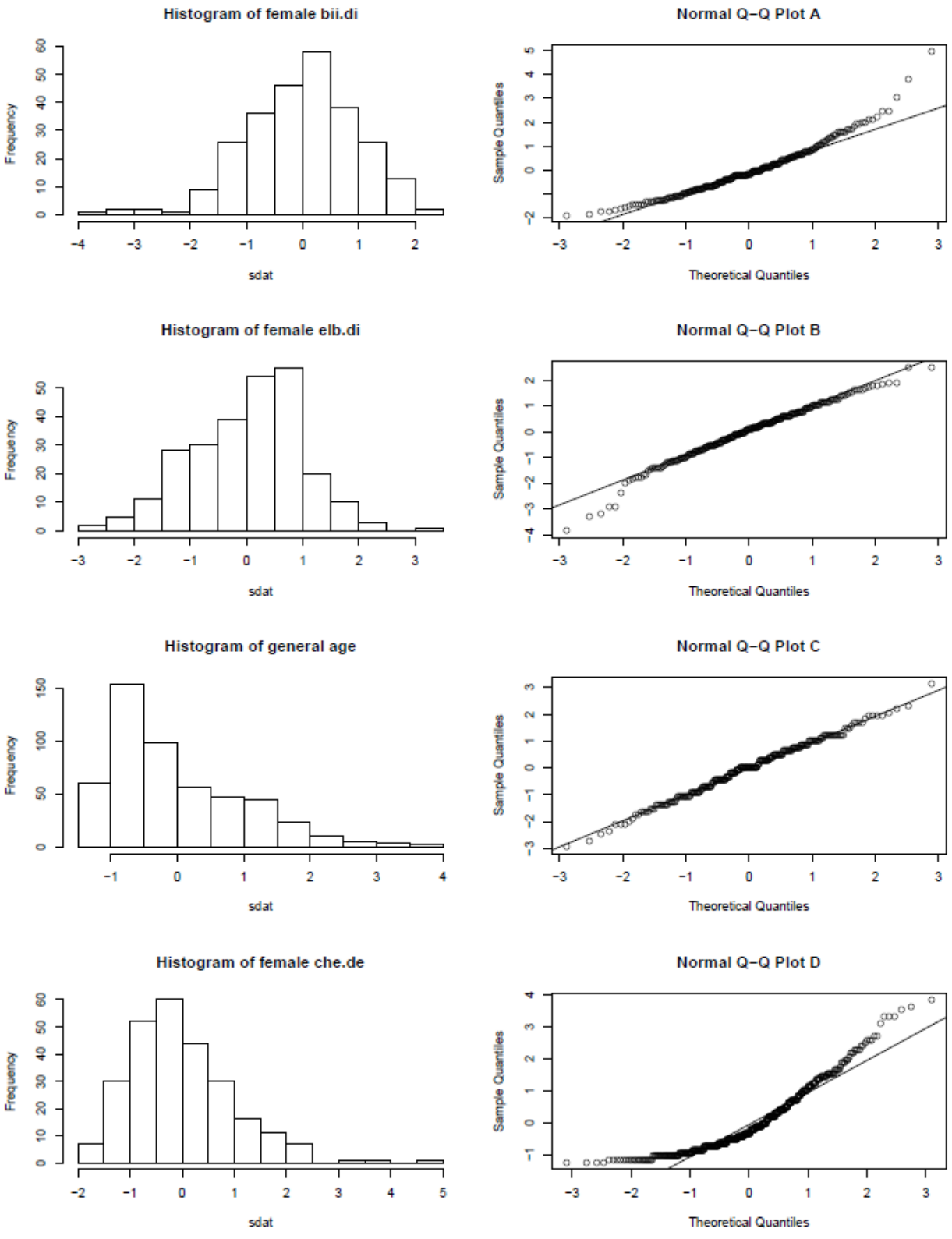


Figure 1: histQQmatch