Exercise3 : Distribution

Shikhar Dhwaj

27 September 2017

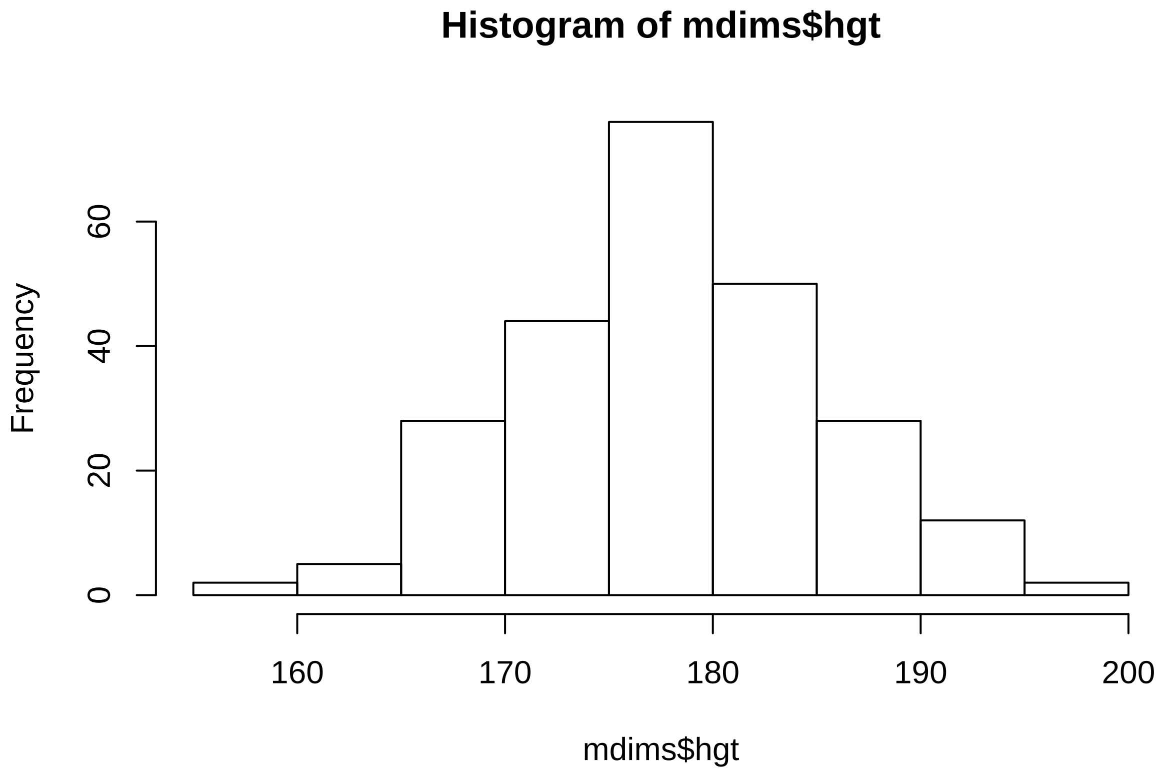
Reading the data file

```
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
```

Q-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?

```
mdims <- bdims[bdims$sex==1,]
fdims <- bdims[bdims$sex==0,]
hist(mdims$hgt)
```



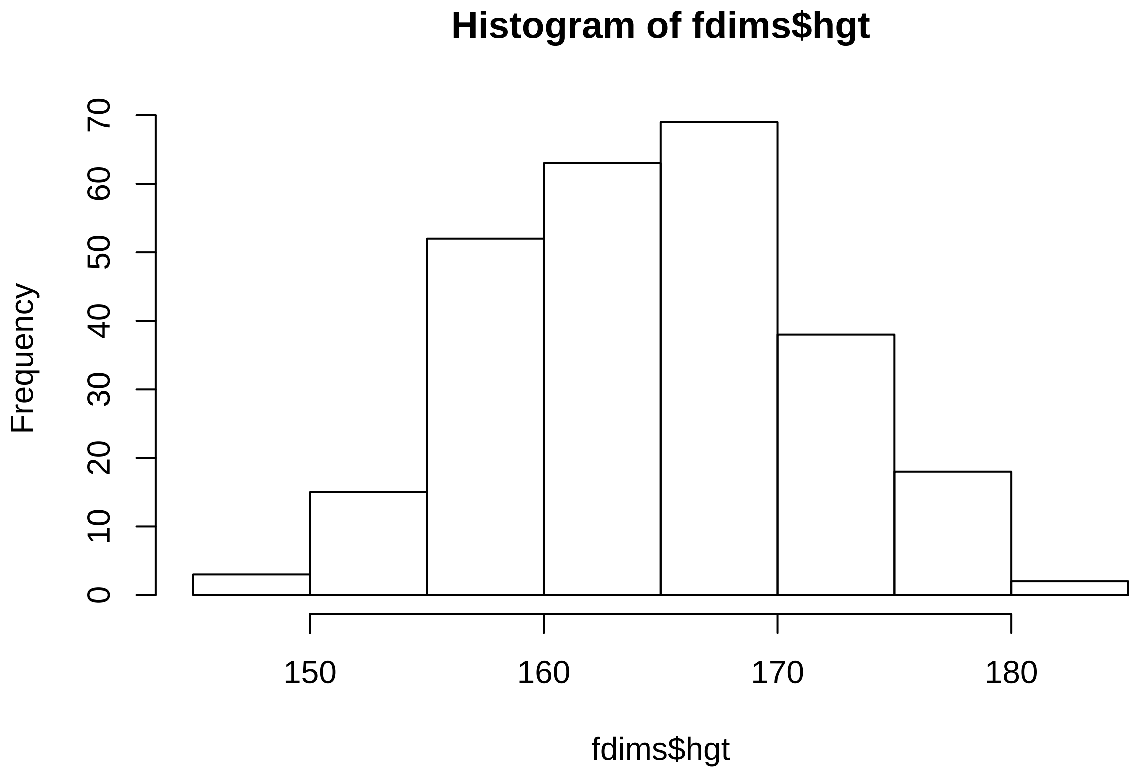
```
cat("Mean = ",mean(mdims$hgt))
```

```
## Mean =  177.7453
```

```
cat("Standard Deviation = ",sd(mdims$hgt))
```

```
## Standard Deviation =  7.183629
```

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



```
cat("Mean = ",mean(fdims$hgt))
```

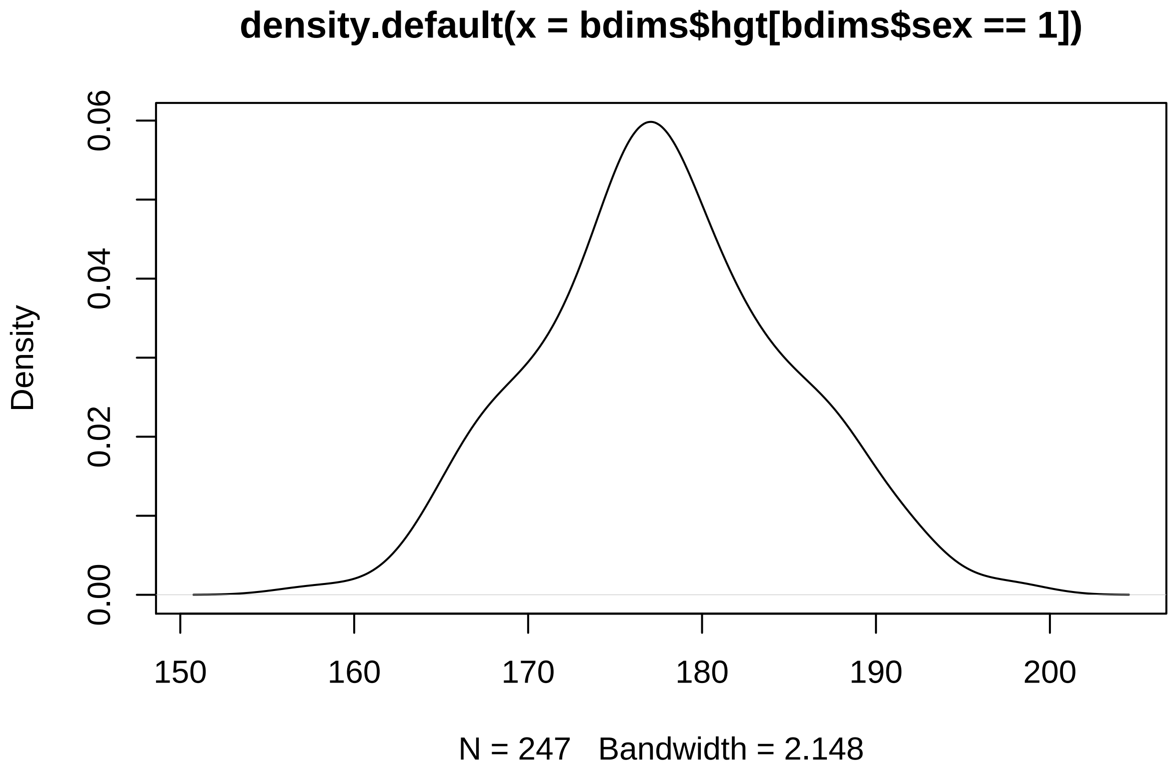
```
## Mean =  164.8723
```

```
cat("Standard Deviation = ",sd(fdims$hgt))
```

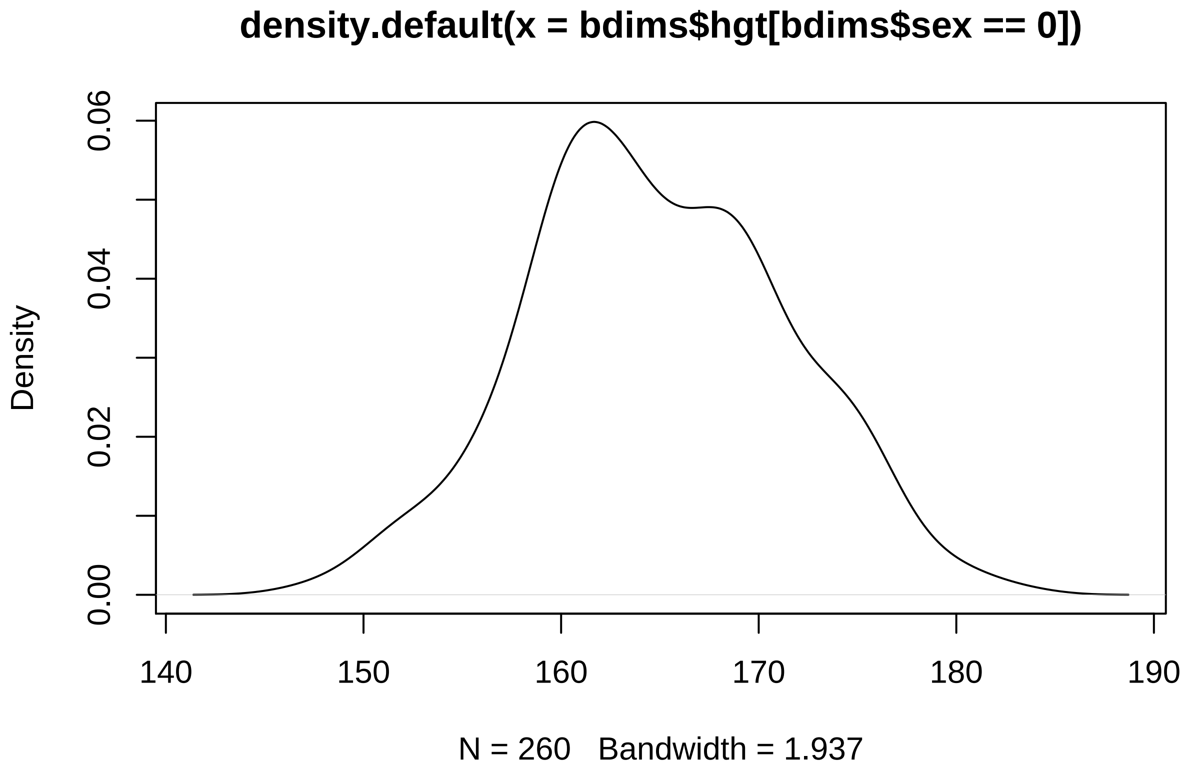
```
## Standard Deviation =  6.544602
```

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

```
plot(density(bdims$hgt[bdims$sex==1]))
```



```r
plot(density(bdims$hgt[bdims$sex==0]))
```

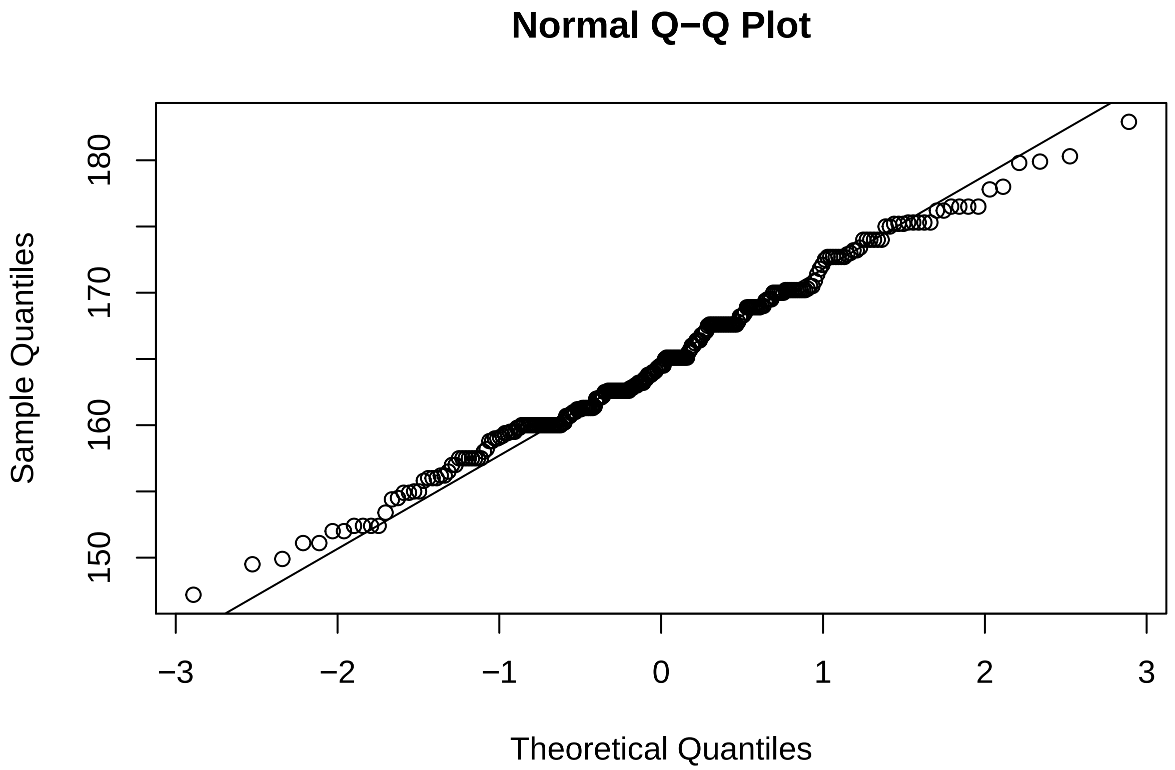


Yes,distribution of heights is nearly normal

Q-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?

Real Data

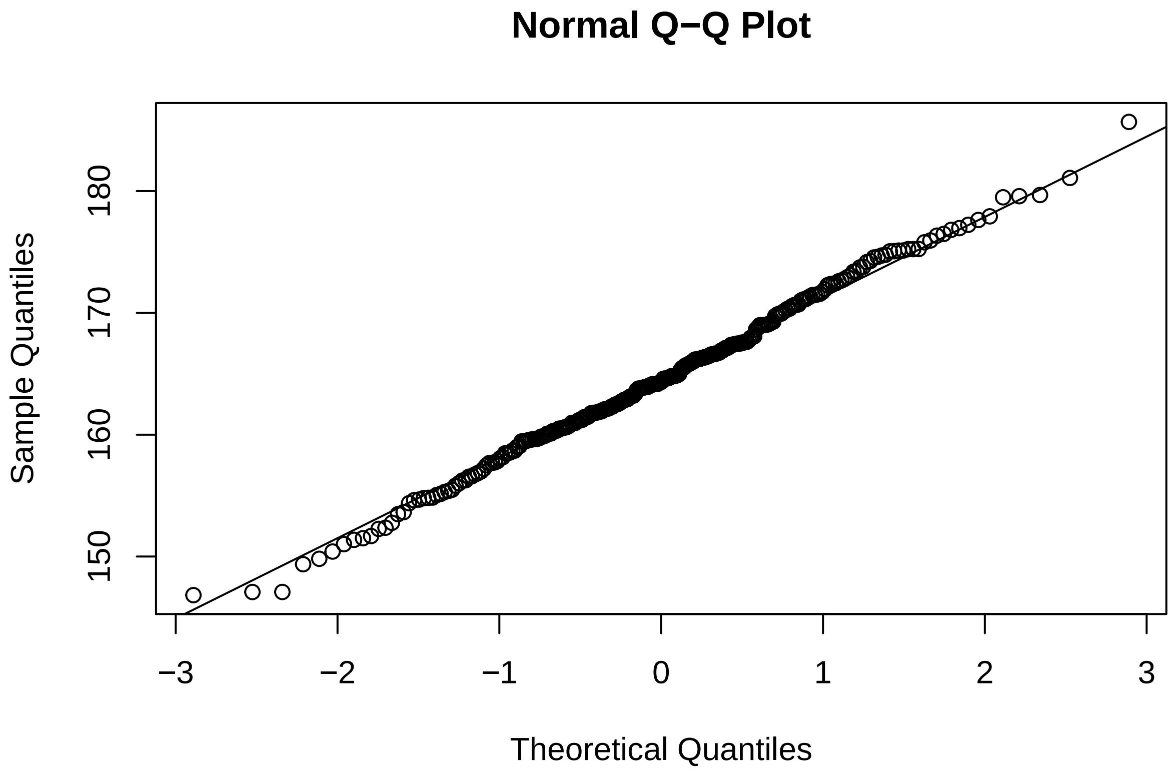
```
qqnorm(fdims$hgt)
qqline(fdims$hgt)
```



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

Simulated Data

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



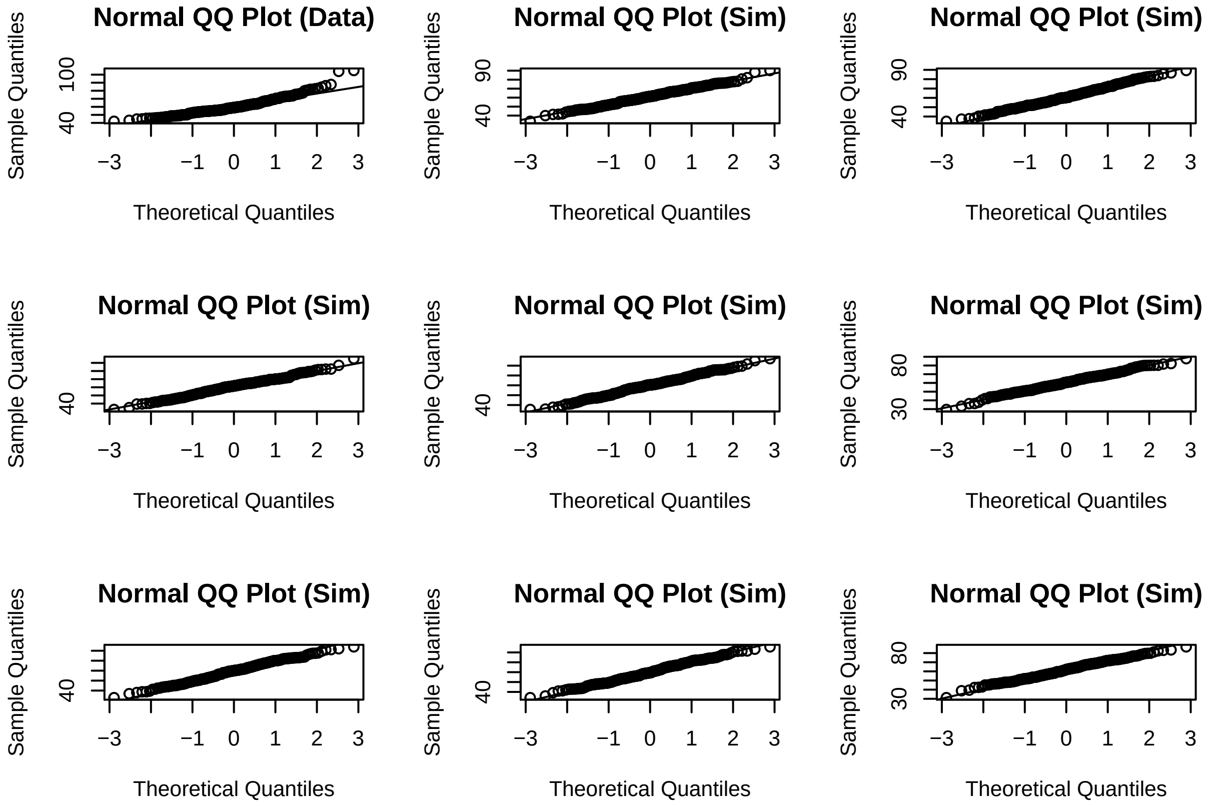
The points in the data don't fall exactly on a line, but they're quite close. The largest deviations come in the tail of the distribution. The simulated plot is more smoothly linear than the data, but that is likely due to a discretization in the data

Q-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 qq plot for female heights is strikingly similar to that from simulated normal data sets. In fact, several of the simulated plots show a greated deviation from linearity in the tails that does the original data. Again, the main difference is that the data has a stairstep shape

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

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



The normal approximation appears to be less appropriate for wgt than for hgt. This data shows some curvature in the shape of the qqplot that suggests a longer right tail that we'd expect from nearly normal data and also shows two notable outliers

Q-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?

If we assume that female heights are normally distributed, we can find this probability by calculating a Z score and consulting a Z table

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

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

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

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

Difference between two probabilities is 0.0005882331 If we assume that female weights are normally distributed, we can find this probability by calculating a Z score and consulting a Z table

```r
1 - pnorm(q = 70, mean = mean(fdims$wgt), sd = sd(fdims$wgt))
```

```
## [1] 0.1641539
```

```r
sum(fdims$wgt > 70) / length(fdims$wgt)
```

```
## [1] 0.1576923
```

Difference between two probabilities is 0.0064616

Q-7)Now let's consider some of the other variables in the body dimensions data set. Using the figures at the end of the exercises, generate the histogram.

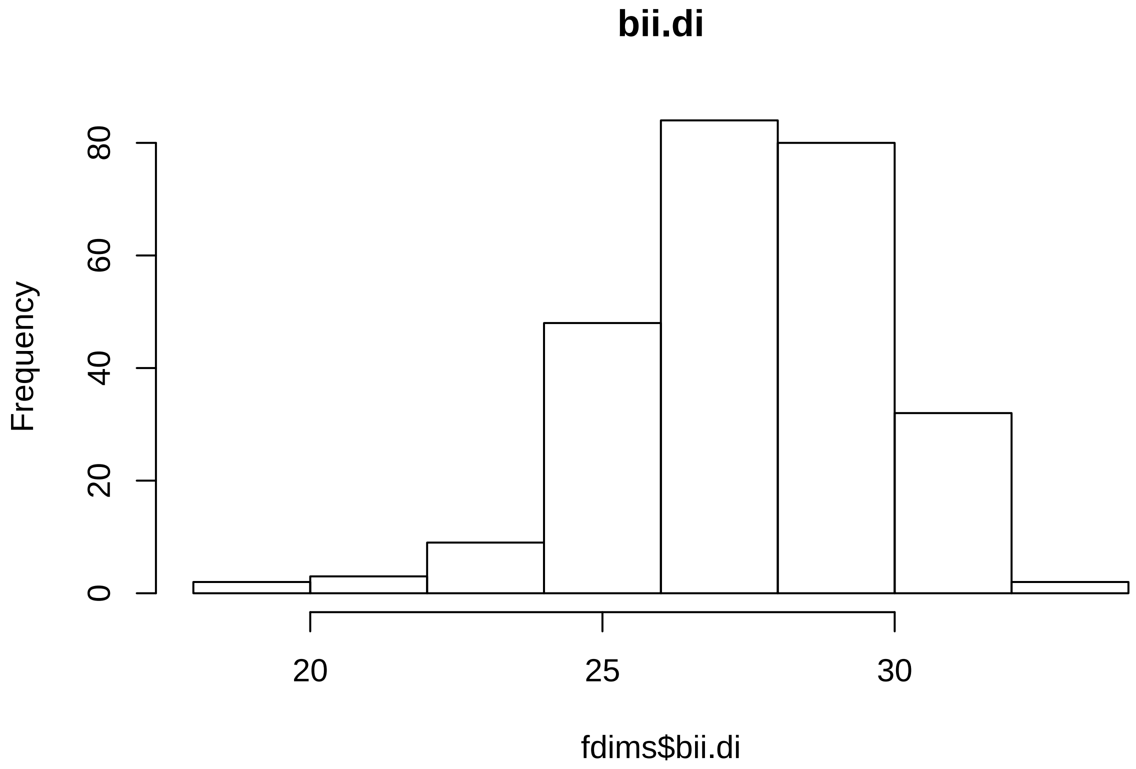
a. The histogram for female biiliac (pelvic) diameter (bii.di)

b. The histogram for female elbow diameter (elb.di)

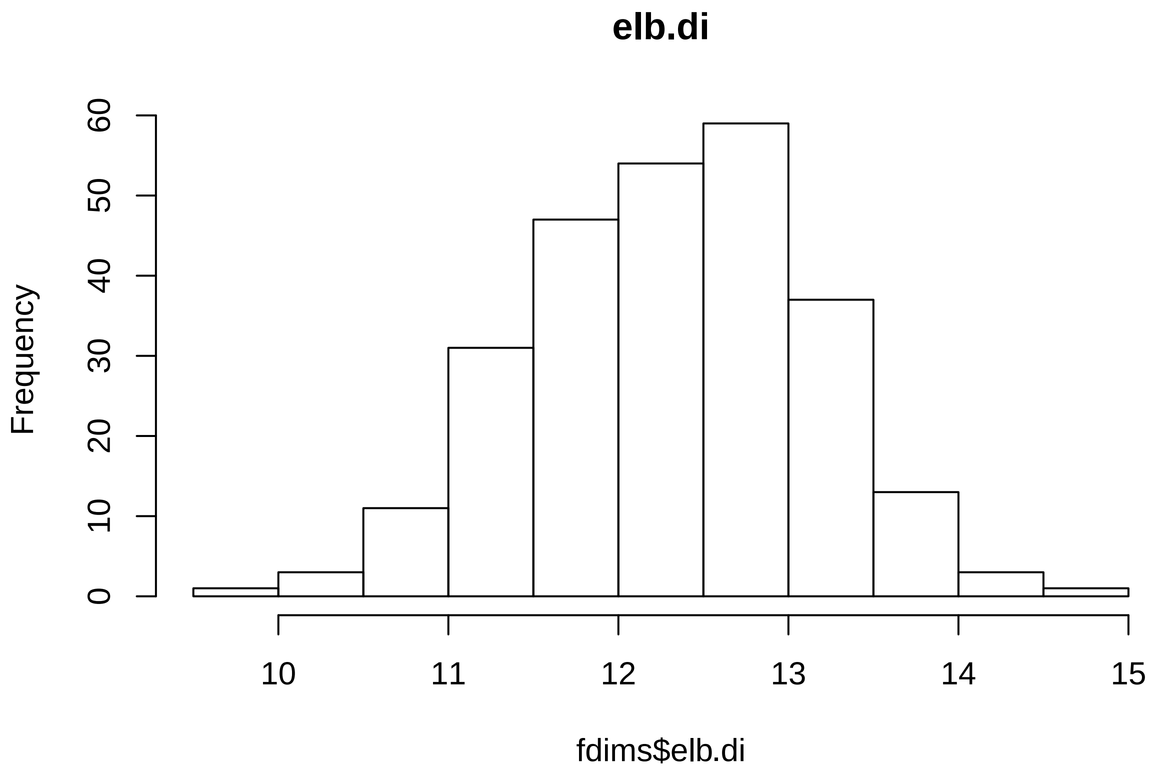
c. The histogram for general age (age)

d. The histogram for female chest depth (che.de)

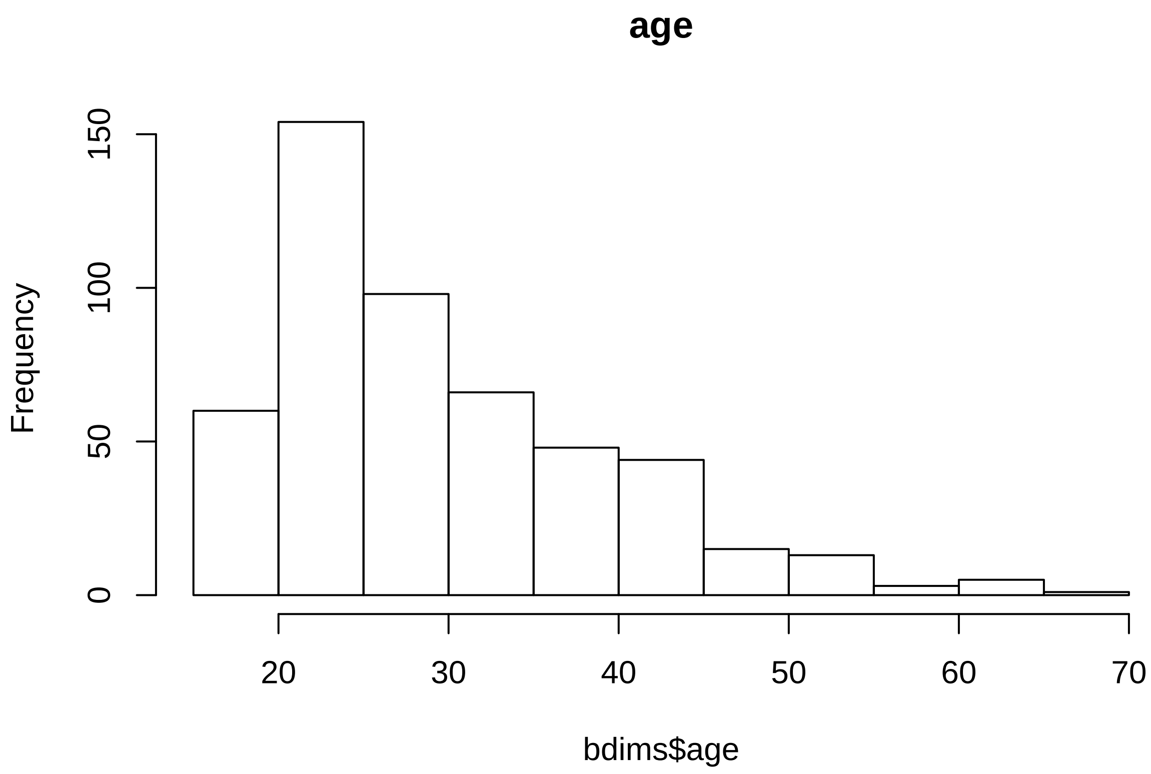
```r
hist(fdims$bii.di,main="bii.di")
```



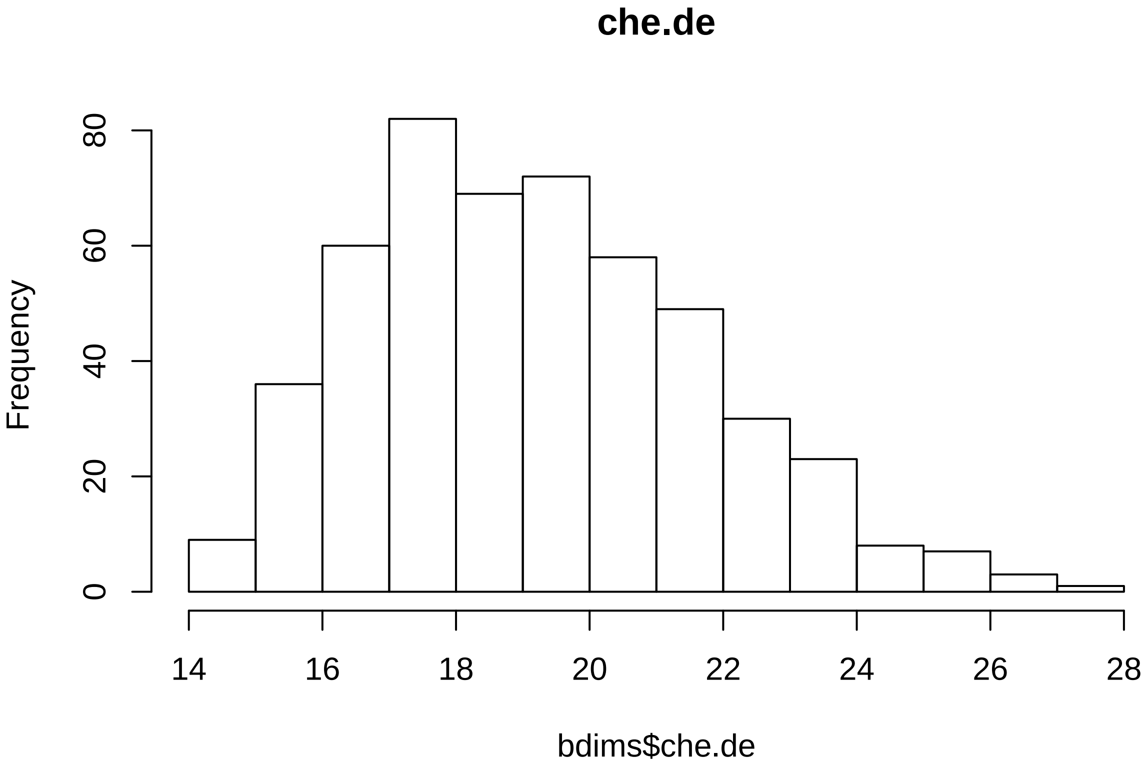
```r
hist(fdims$elb.di,main="elb.di")
```



```r
hist(bdims$age,main="age")
```



```
hist(bdims$che.de,main="che.de")
```



Q-8)Note that normal probability plots C and D have a slight stepwise pattern.Why do you think this is the case?

This is likely due to the discrete scale on which the data was measured. When people report there age, they usually only provide integer values, not ages like 28.3746 years. This is what creates the step patter in the variable on the y-axis of the qqplot. The x-axis refers to the percentiles of the normal distribution, which is continuous, so the plots are continuous in their x-values.