# Module 04 Lab

Daniel Jackson

September 19th, 2023

## Lab Data

```
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 che.gi
## 1   42.9   26.0   31.5   17.7   28.0   13.1   10.4   18.8   14.1  106.2   89.5
## 2   43.7   28.5   33.5   16.9   30.8   14.0   11.8   20.6   15.1  110.5   97.0
## 3   40.1   28.2   33.3   20.9   31.7   13.9   10.9   19.7   14.1  115.1   97.5
## 4   44.3   29.9   34.0   18.4   28.2   13.9   11.2   20.9   15.0  104.5   97.0
## 5   42.5   29.9   34.0   21.5   29.4   15.2   11.6   20.7   14.9  107.5   97.5
## 6   43.3   27.0   31.5   19.6   31.3   14.0   11.5   18.8   13.9  119.8   99.9
##   wai.gi nav.gi hip.gi thi.gi bic.gi for.gi kne.gi cal.gi ank.gi wri.gi age
## 1   71.5   74.5   93.5   51.5   32.5   26.0   34.5   36.5   23.5   16.5  21
## 2   79.0   86.5   94.8   51.5   34.4   28.0   36.5   37.5   24.5   17.0  23
## 3   83.2   82.9   95.0   57.3   33.4   28.8   37.0   37.3   21.9   16.9  28
## 4   77.8   78.8   94.0   53.0   31.0   26.2   37.0   34.8   23.0   16.6  23
## 5   80.0   82.5   98.5   55.4   32.0   28.4   37.7   38.6   24.4   18.0  22
## 6   82.5   80.1   95.3   57.5   33.0   28.0   36.6   36.1   23.5   16.9  21
##    wgt   hgt sex
## 1 65.6 174.0   1
## 2 71.8 175.3   1
## 3 80.7 193.5   1
## 4 72.6 186.5   1
## 5 78.8 187.2   1
## 6 74.8 181.5   1
```

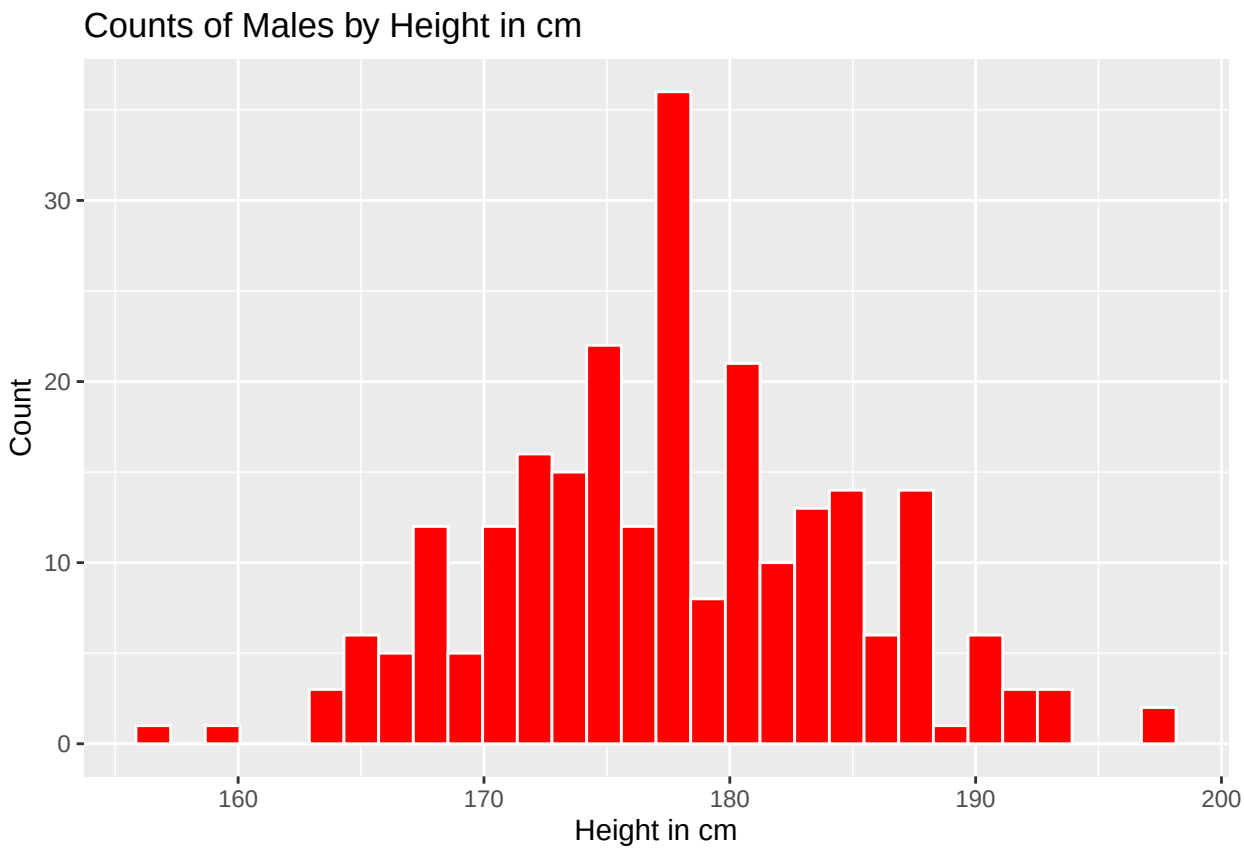
```
mdims <- subset(bdims, sex == 1)
fdims <- subset(bdims, sex == 0)
```

## Exercise 1

```
library(ggplot2)
# Histogram for males
```

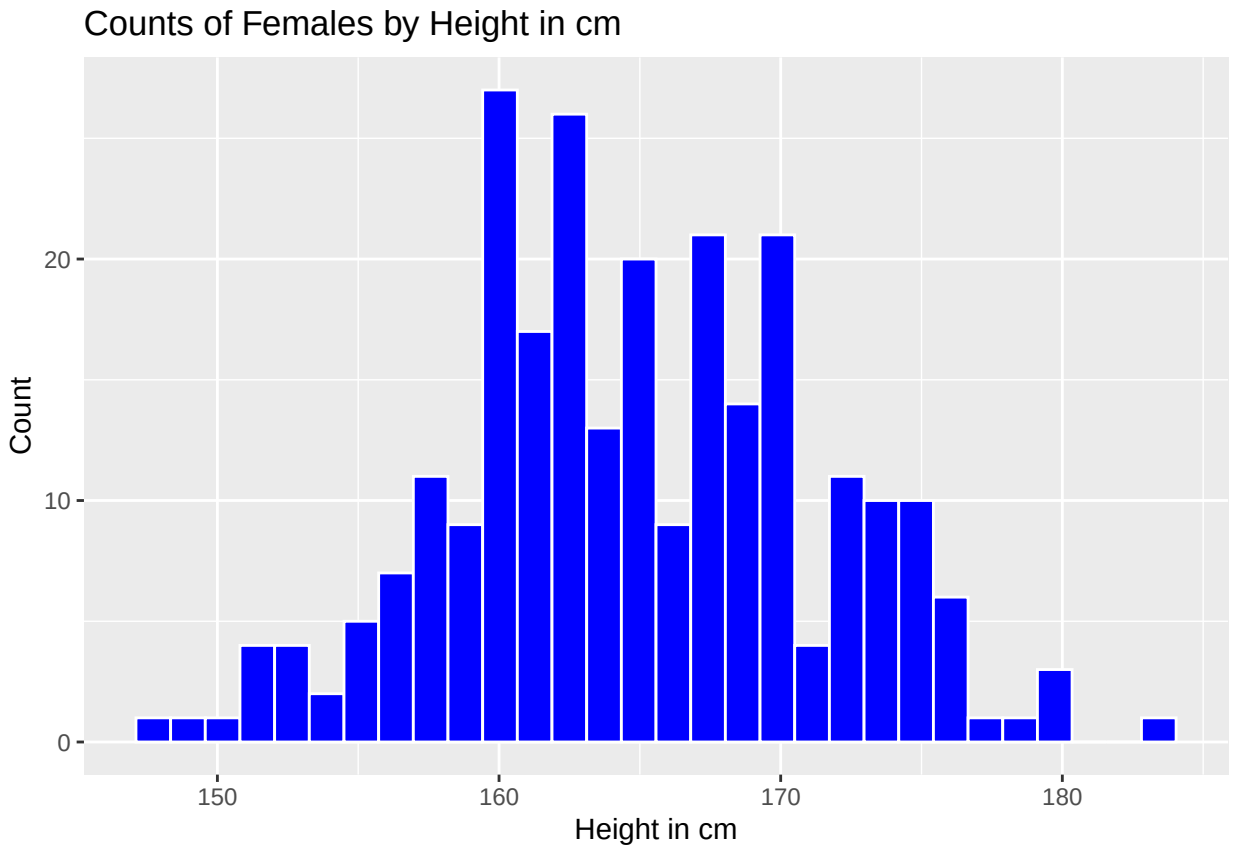
```
ggplot(mdims,
       mapping = aes(x = hgt)) +
  geom_histogram(color = 'white', fill = 'red') +
  labs(x = 'Height in cm',
       y = 'Count',
       title = 'Counts of Males by Height in cm')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



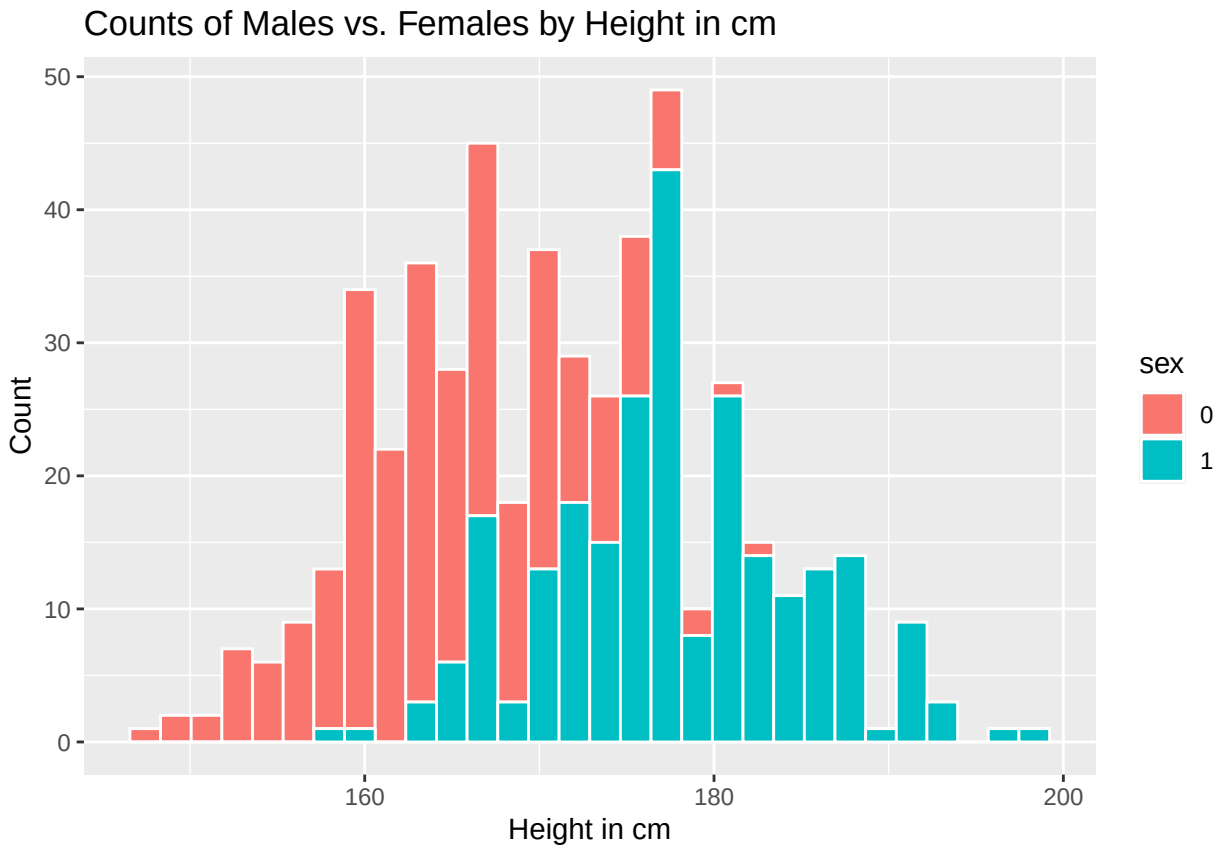

```
# Histograms for females
ggplot(fdims,
       mapping = aes(x = hgt)) +
  geom_histogram(color = 'white', fill = 'blue') +
  labs(x = 'Height in cm',
       y = 'Count',
       title = 'Counts of Females by Height in cm')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# One histogram with height by males and females
ggplot(bdims,
       mapping = aes(x = hgt, fill = sex)) +
  geom_histogram(color = 'white') +
  labs(x = 'Height in cm',
       y = 'Count',
       title = 'Counts of Males vs. Females by Height in cm')
```

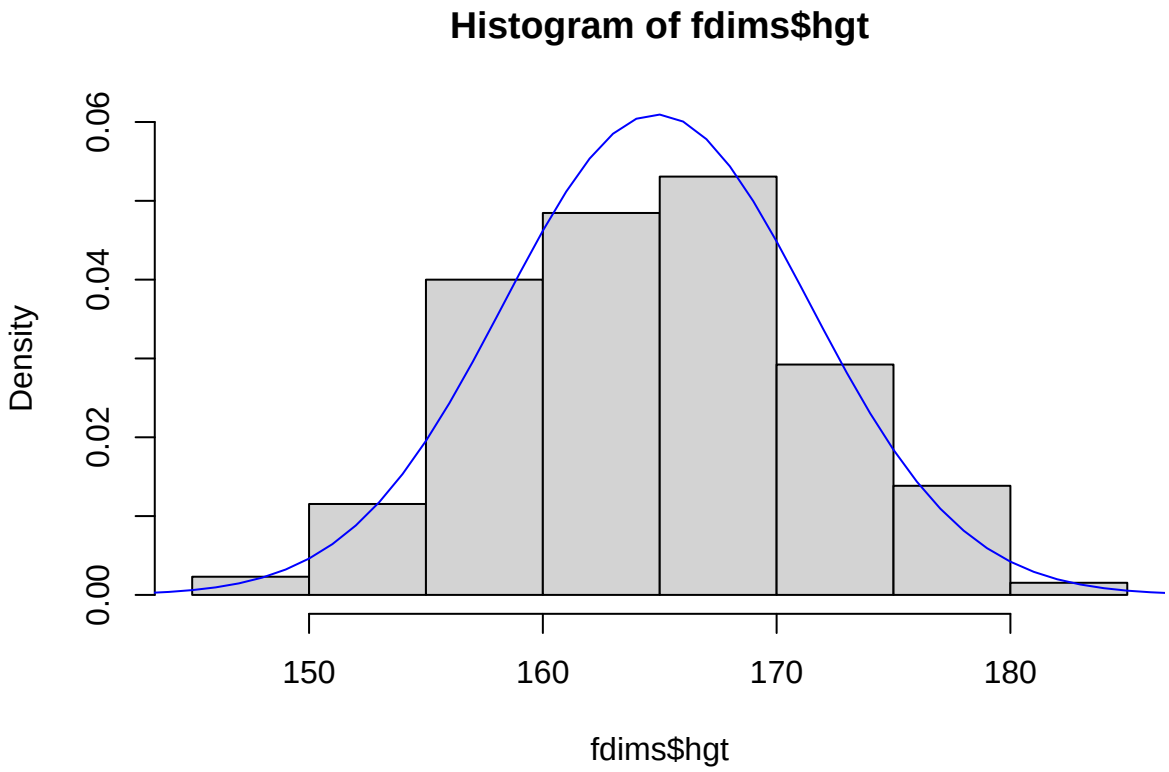
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




Both plots look relatively normal. Both distributions are unimodial and the males histogram looks almost symmetric. Both have a bell-shape to them.

## Exercise 2

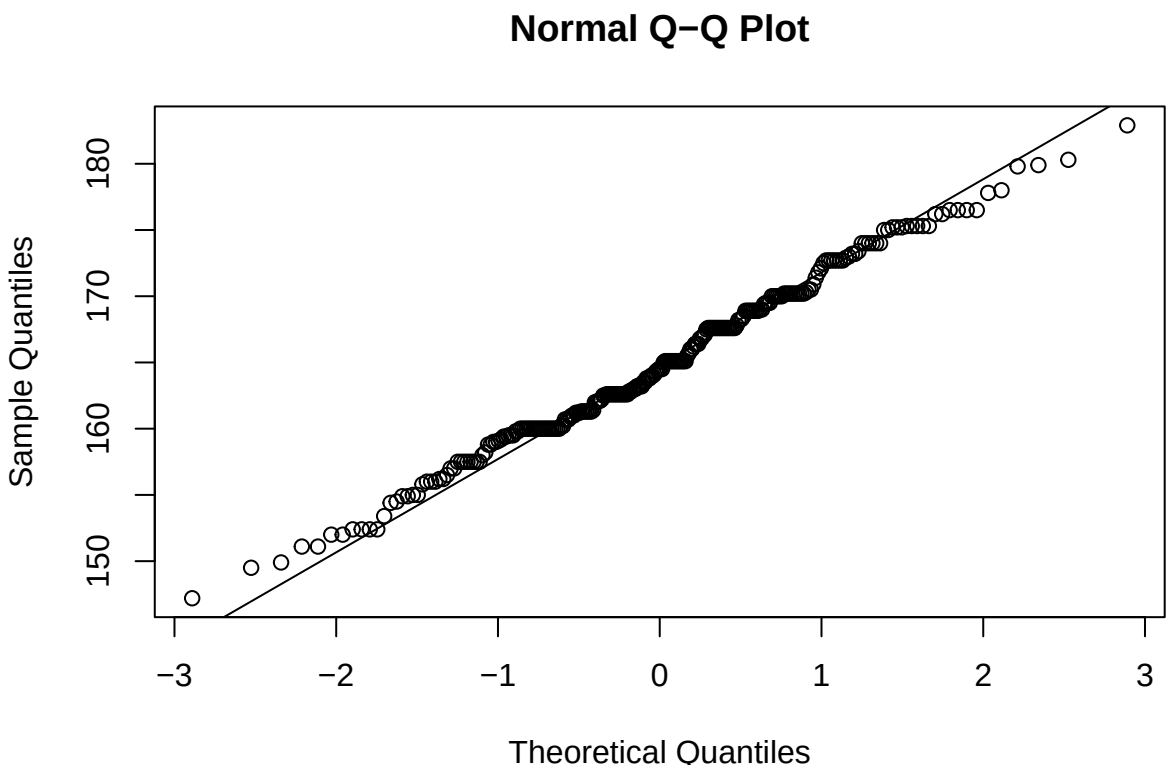
```
# Working with woman's height, find mean and sd:
fhgtmean <- mean(fdims$hgt)
fhgtsd   <- sd(fdims$hgt)

# Make density histogram to use as backdrop and overlay a normal probability curve
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")
```



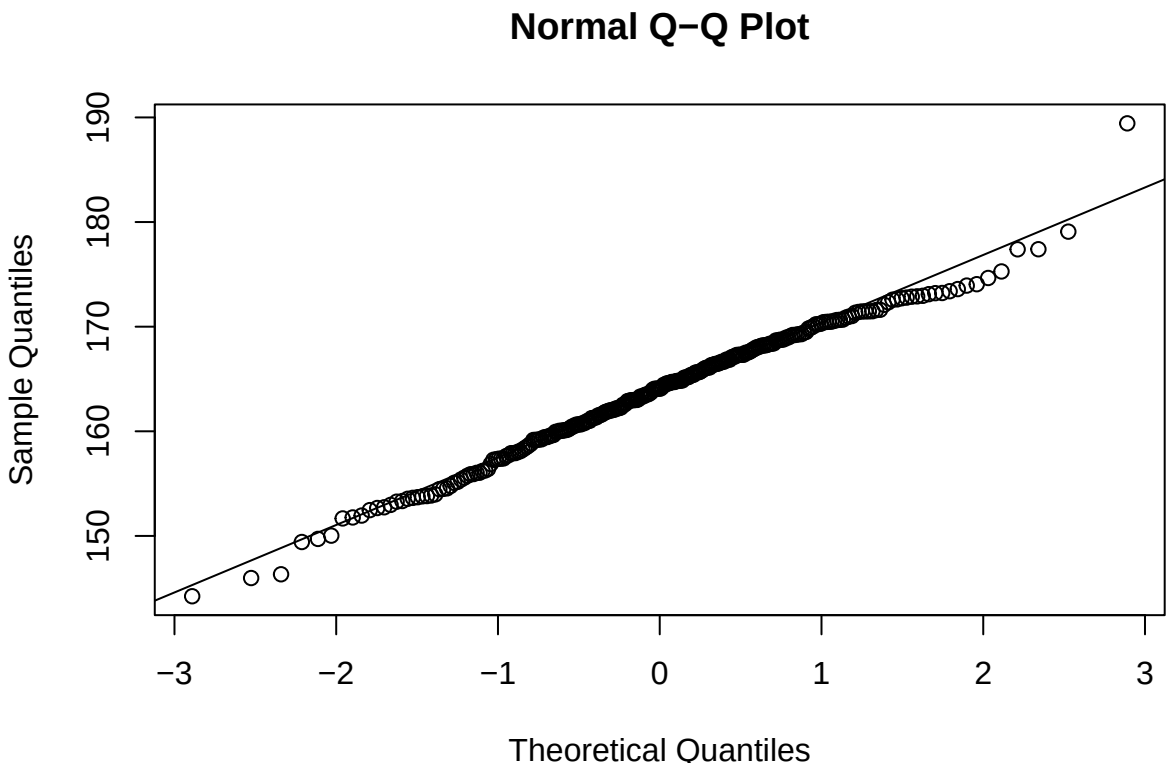

Based on this density histogram, I would say the data follows a nearly normal distribution.

## Exercise 3

```
# We can also construct a normal probability plot to check if distribution is nearly normal. This plot
## follow the line.

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




```
# What do probability plots look like for data that I know came from a normal distribution? We can answ

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

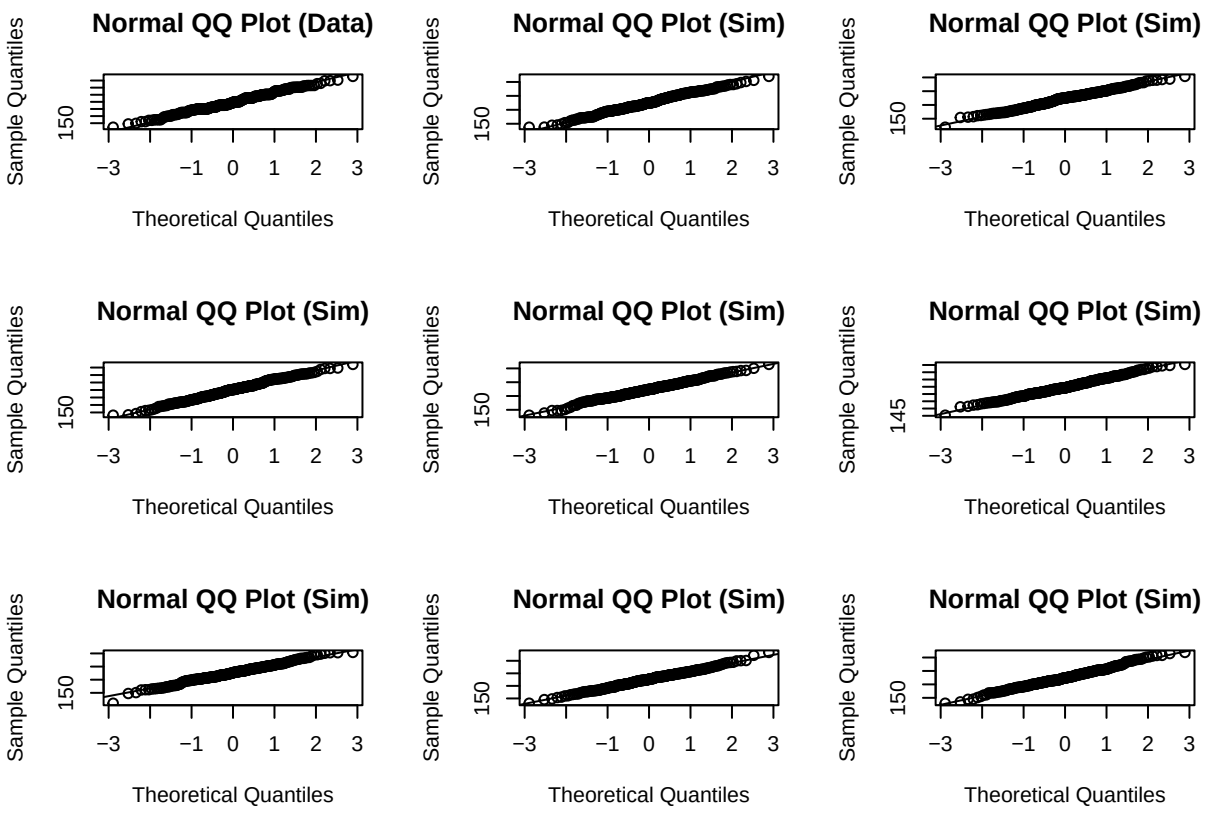


Not all points fall on the line. But it seems that more points are closer to the line compared to the probability plot for the real data.The points near the line seem to be more linear in the probability plot for sim_norm compared to the real data.

## Exercise 4

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.

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



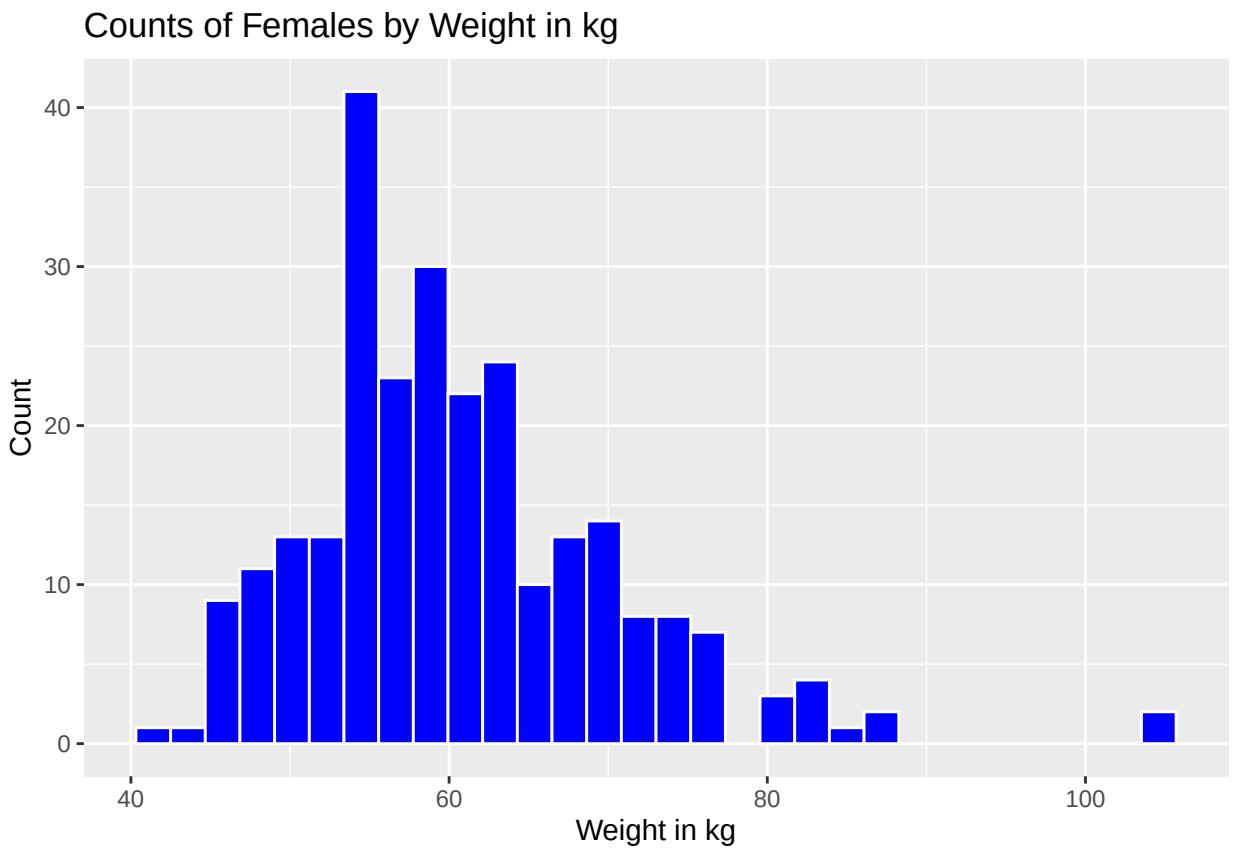

All of the outputs look relatively linear, and very similar to real data.

## Exercise 5

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

```
ggplot(fdims,
       mapping = aes(x = wgt)) +
  geom_histogram(color = 'white', fill = 'blue') +
  labs(x = 'Weight in kg',
       y = 'Count',
       title = 'Counts of Females by Weight in kg')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



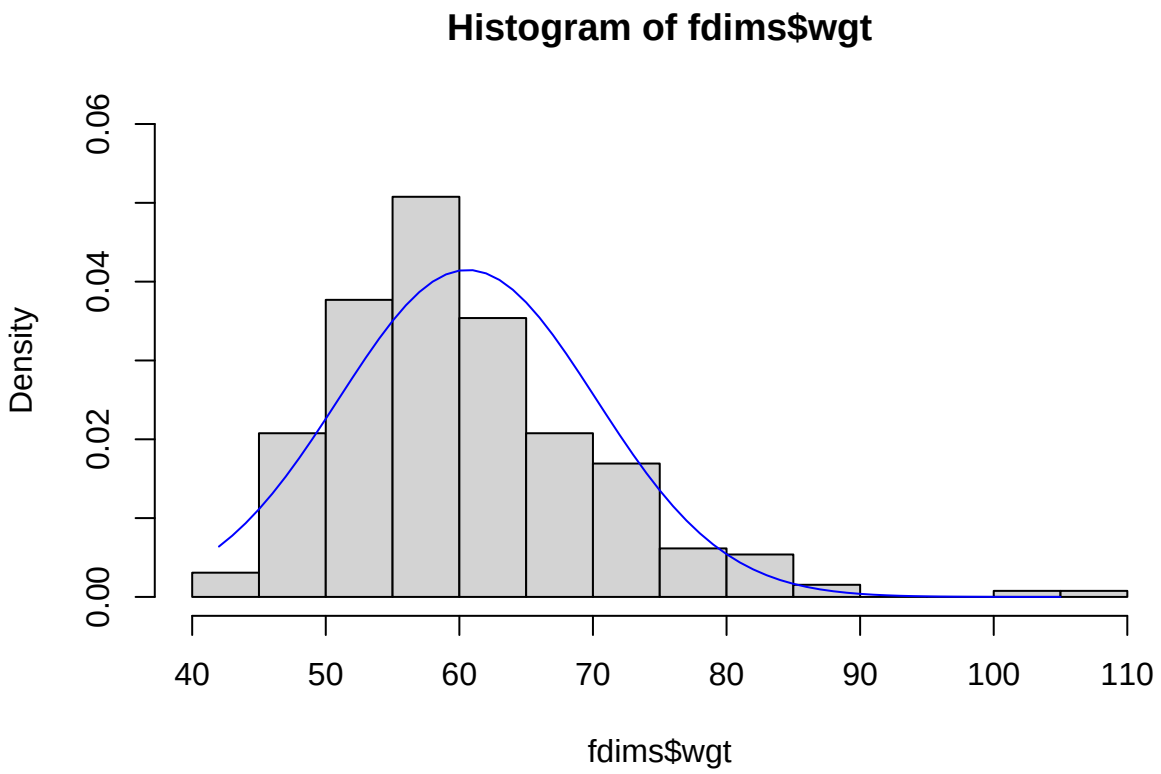

```
# Working with woman's weight, find mean and sd:
fwgtmean <- mean(fdims$wgt)
fwgtsd   <- sd(fdims$wgt)


# Make density histogram to use as backdrop and overlay a normal probability curve
min(fdims$wgt)
```

```
## [1] 42
```

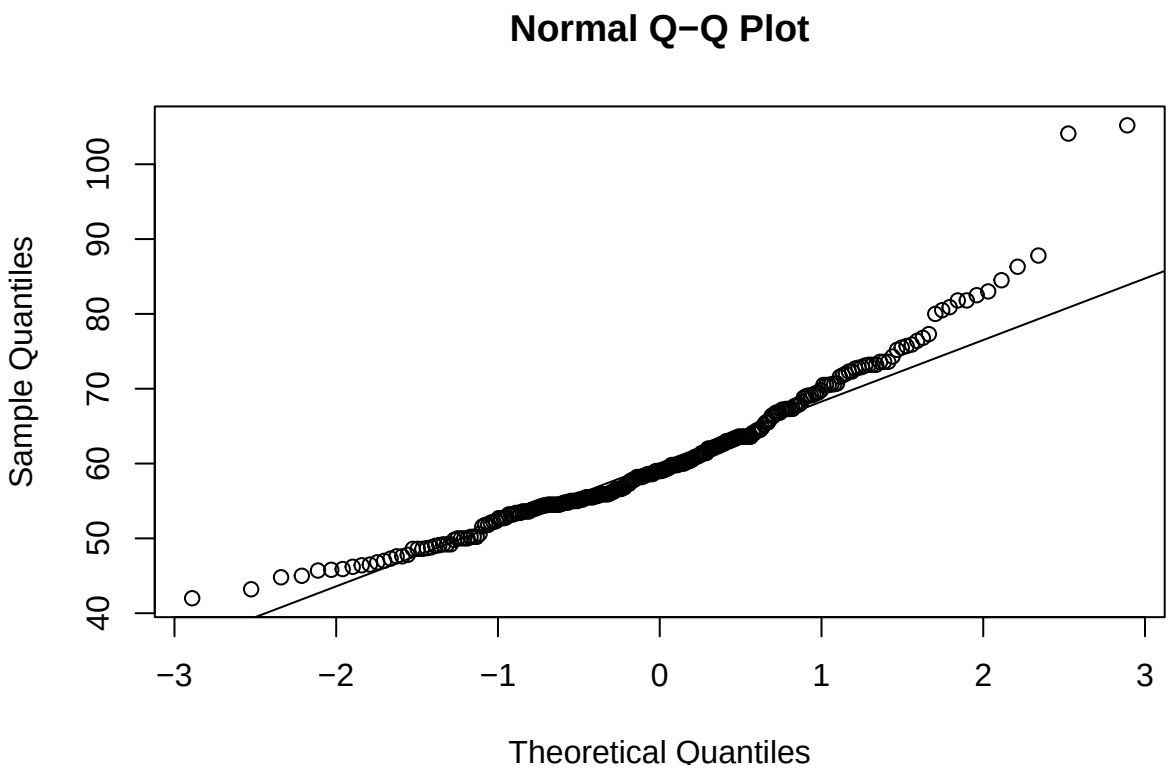
```
max(fdims$wgt)
```

```
## [1] 105.2
```

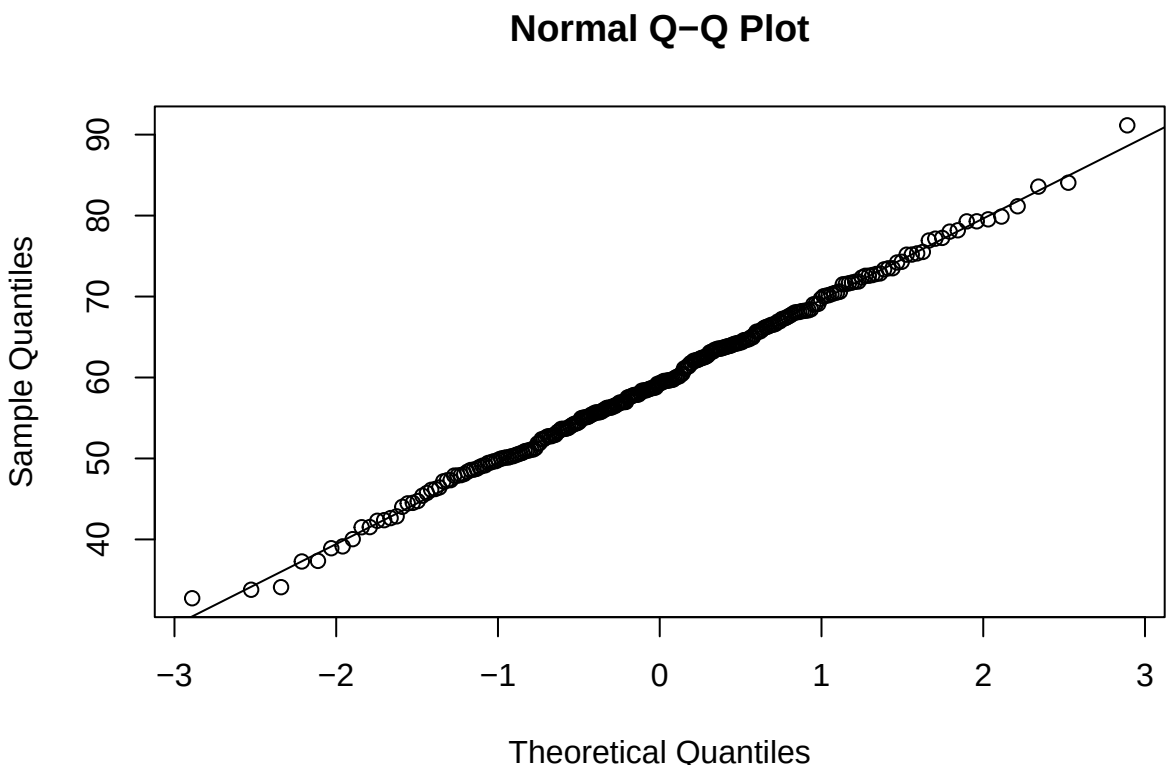
```
hist(fdims$wgt, probability = TRUE, ylim = c(0, 0.06))
x <- 42:105.2
y <- dnorm(x = x, mean = fwgtmean, sd = fwgtsd)
lines(x = x, y = y, col = "blue")
```




```
# Based on the distribution from the density histogram, it does not look like the
# distribution is normal. Looks more right-skewed than symmetric.

# Let's look at normal probability plot:

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




```
# The probablity plot is concave up which further confirms my thoughts that the
# distribution is right-skewed.

# Let's simulate data from a normal distribution using rnorm.
sim_norm_wgt <- rnorm(n = length(fdims$wgt), mean = fwgtmean, sd = fwgtsd)
qqnorm(sim_norm_wgt)
qqline(sim_norm_wgt)
```




This data looks s-shaped more than linear. Therefore, I am going to say that the female weight distribution is right-skewed and not normal.

## Exercise 6

What is the probability that a randomly chosen young adult female is taller than 6 feet (about 182 cm)?

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

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

```
# Returned 0.0044 which is 0.44%

# Assuming a normal distribution has allowed us to calculate a theoretical probability.If we want to ca

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

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

```
# This returned 0.0038 which is 0.38%
```

Probability that a randomly chose young adult femaile is taller than 6 feet is 0.44%

Question 1: What is the probability that a randomly chosen young adult female is smaller than 5 feet 7 inches (about 170 cm)?

```r
print(pnorm(170, mean = fhgtmean, sd = fhgtsd))
```

```
## [1] 0.7833331
```

```r
# Returned 0.7833 which is 78.33%

# Calculate probability empirically
print(sum(fdims$hgt < 170) / length(fdims$hgt))
```

```
## [1] 0.7538462
```

```r
# Returned 0.7538 which is 75.38%

# Difference between two probabilities:
print(78.33 - 75.38)
```

```
## [1] 2.95
```

```r
# Returned 2.95
```

Probability is 78.33%. Probability empirically is 75.38% Difference between the two probabilities is 2.95%

Question 2: What is the probability that a randomly chosen young adult female weighs more than 55 kg?

```r
print(1 - (pnorm(55, mean = fwgtmean, sd = fwgtsd)))
```

```
## [1] 0.7198584
```

```r
# This returned 0.7199 which is 71.99%

# Calculate probability empirically
print(sum(fdims$wgt > 55) / length(fdims$wgt))
```

```
## [1] 0.6923077
```

```r
# This returned 0.6923 which is 69.23%

# Difference between two probabilities:
print(71.99 - 69.23)
```

```
## [1] 2.76
```

```r
# Returned 2.76
```

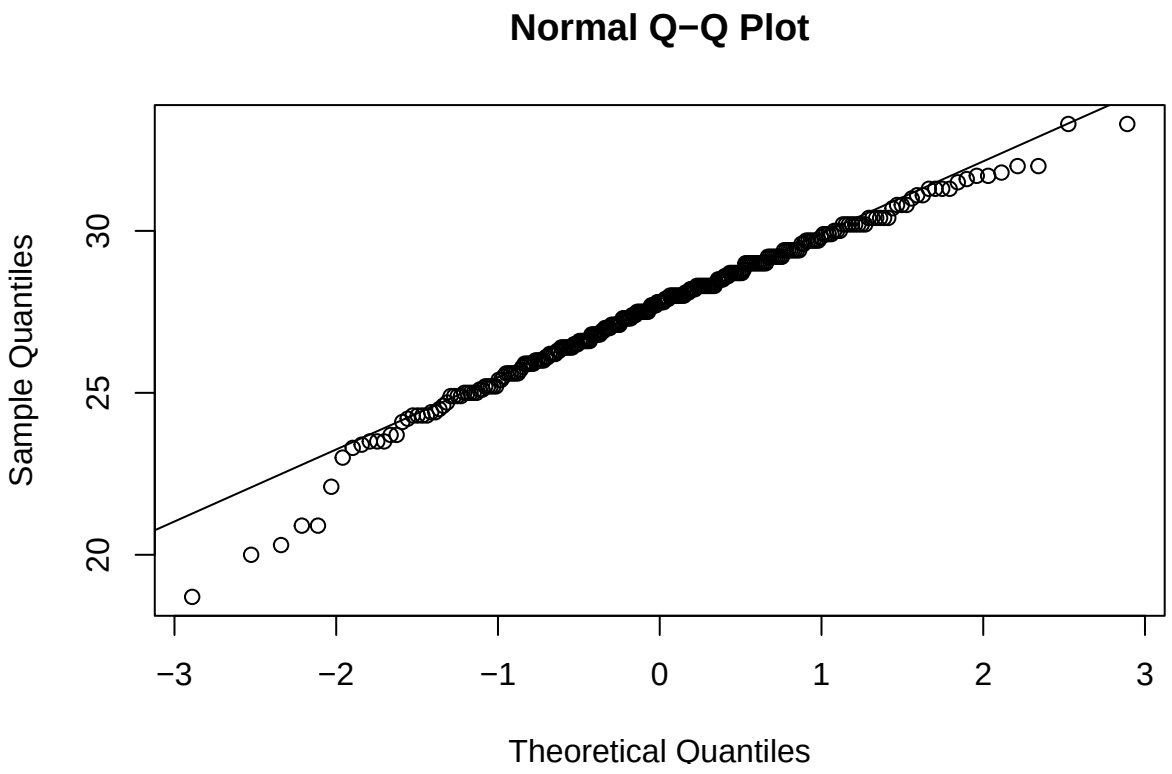
Probability is 71.99%. Probability empirically is 69.23% Difference between the two probabilities is 2.76%

Weight had a closer agreement between two methods. However, height is more normal than weight, so I believe the probabilities of height are more accurate.

## On Your Own Lab

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

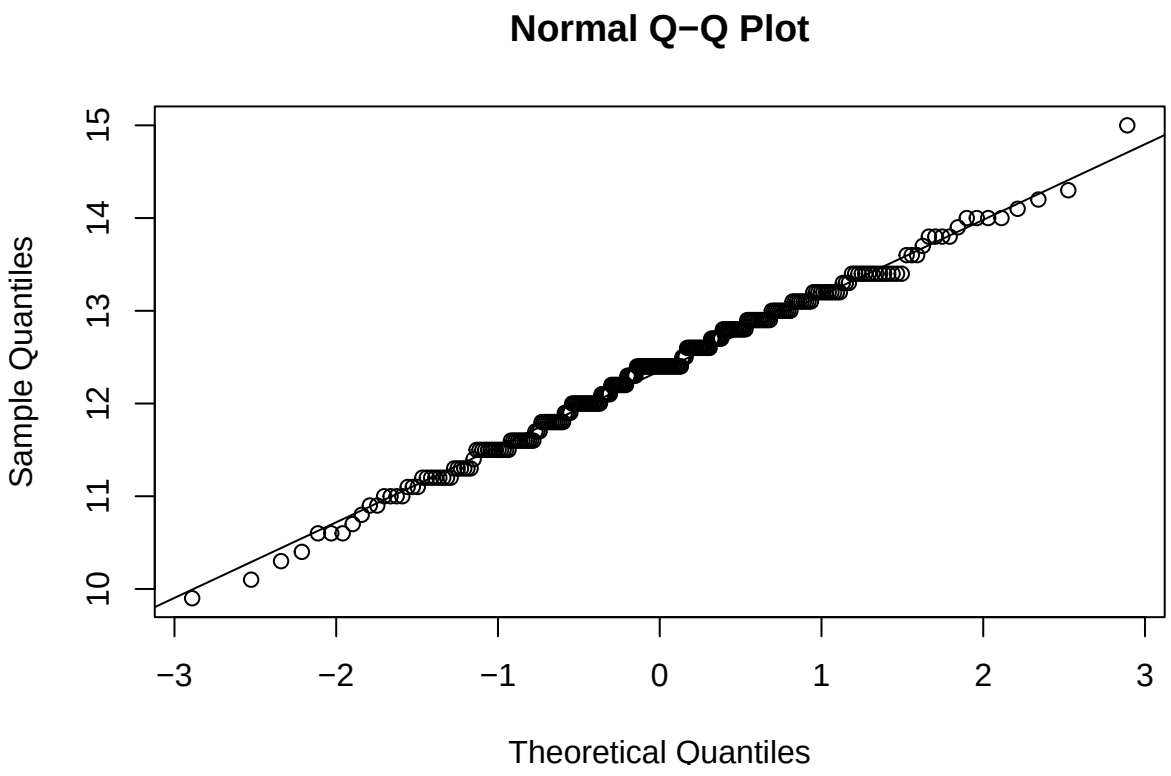
```
qqnorm(fdims$bii.di)
qqline(fdims$bii.di)
```




Belongs to plot B

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

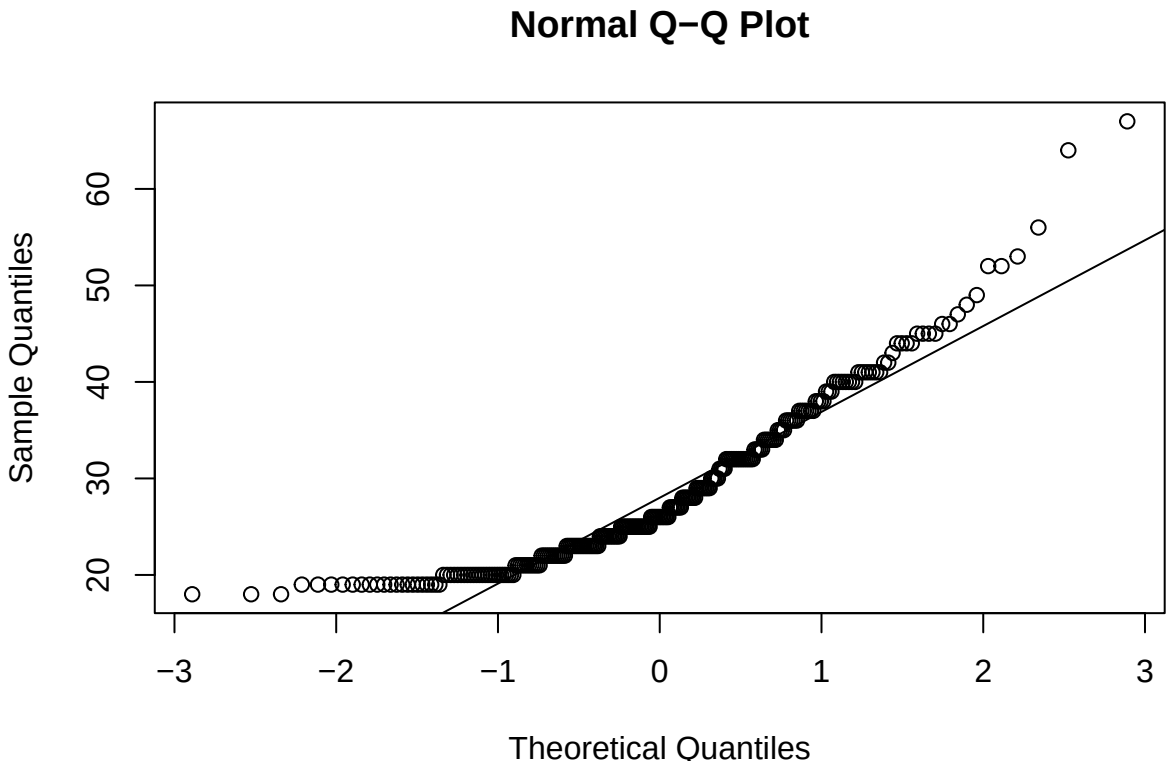
```
qqnorm(fdims$elb.di)
qqline(fdims$elb.di)
```




Belongs to plot C

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

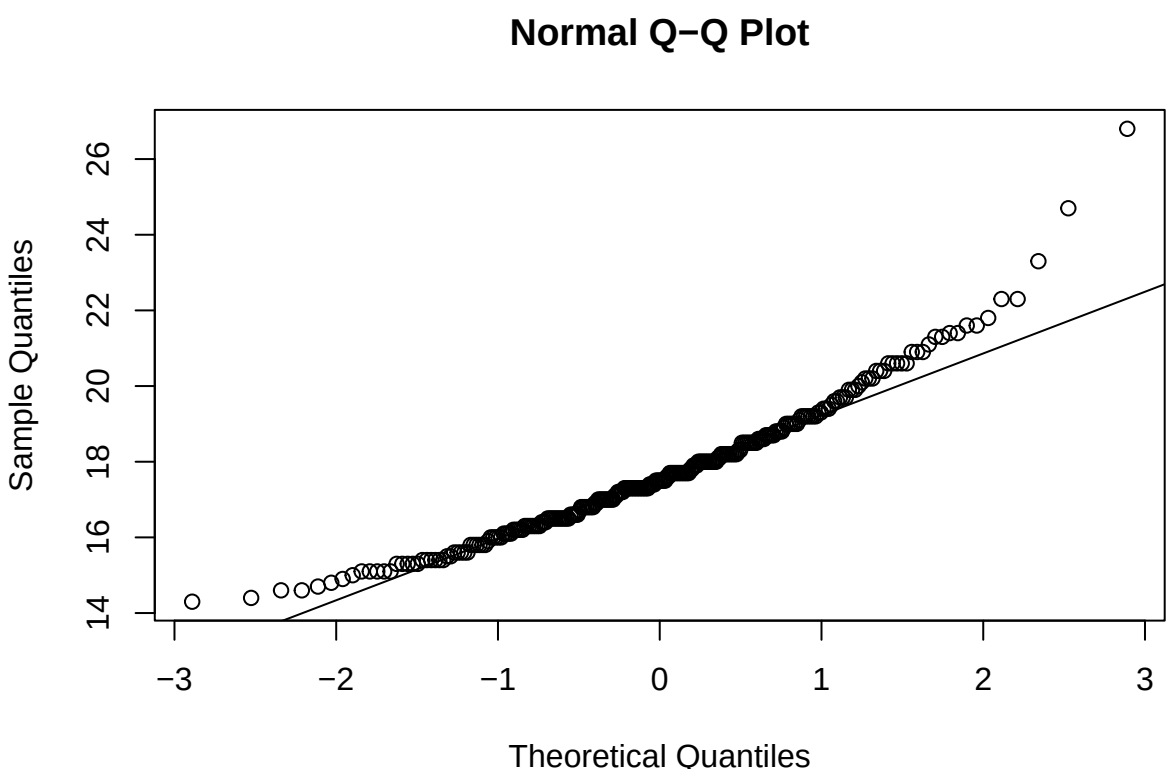
```
qqnorm(fdims$age)
qqline(fdims$age)
```



Belongs to plot D

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

```
qqnorm(fdims$che.de)
qqline(fdims$che.de)
```

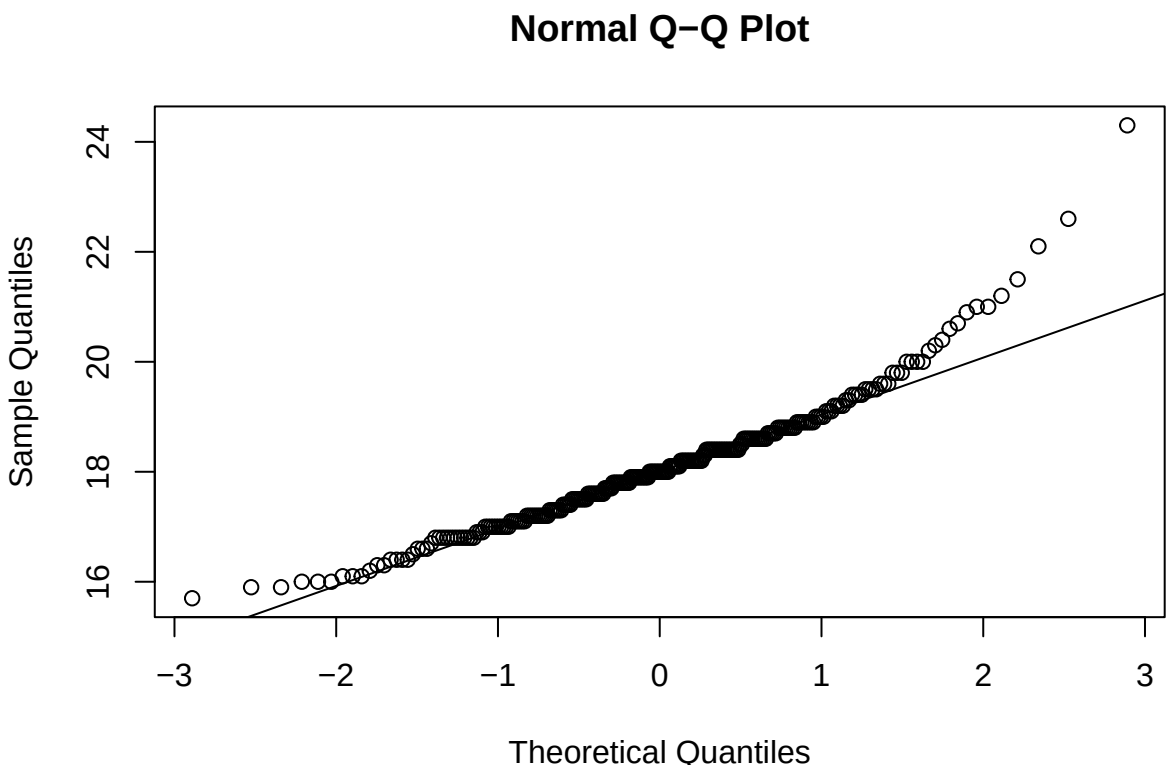



Belongs to plot A

2.) This occurs because these distributions are not normal. The more ouutliers there are, the more skewed the data.

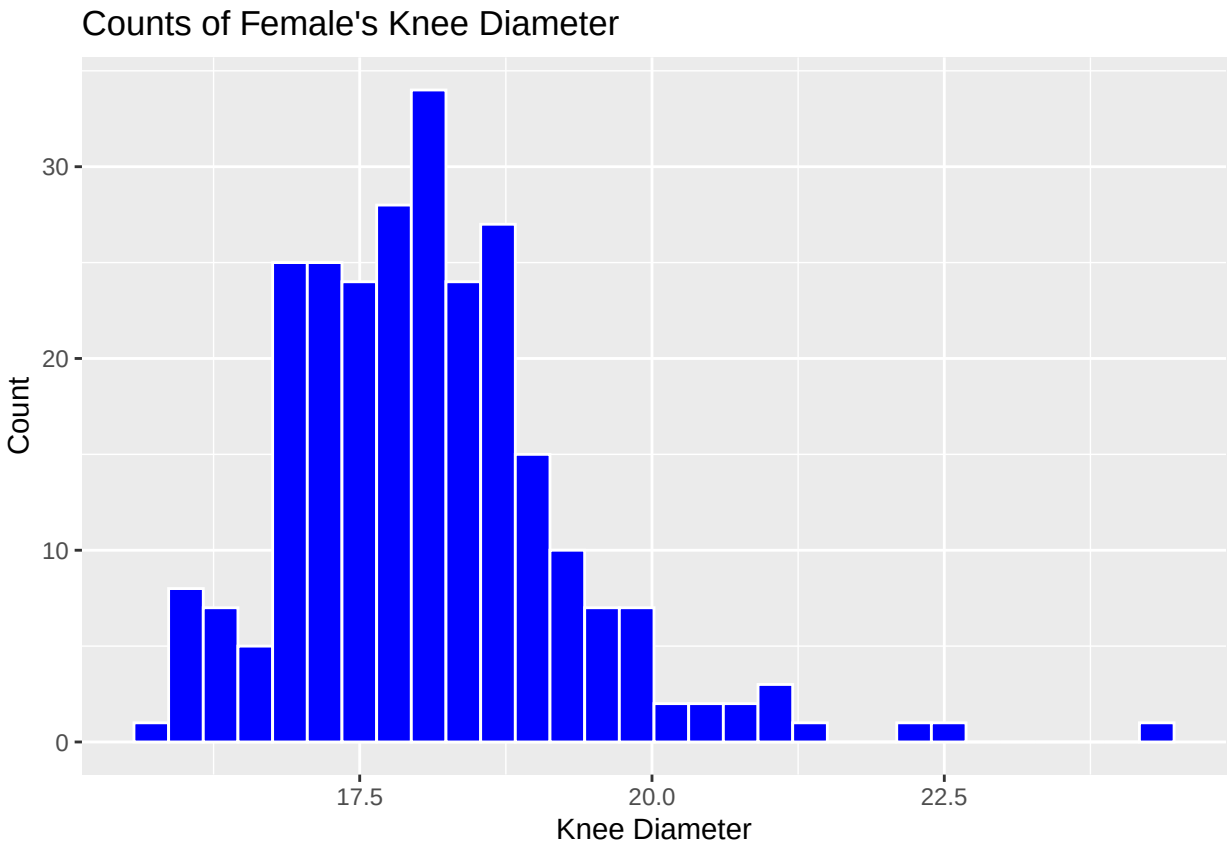
3.)

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




```
# Based on probability plot being concave up, this looks right-skewed.

# Create histogram to confirm conclusion.
ggplot(fdims,
       mapping = aes(x = kne.di)) +
  geom_histogram(color = 'white', fill = 'blue') +
  labs(x = 'Knee Diameter',
       y = 'Count',
       title = "Counts of Female's Knee Diameter")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




Confirmed with histogram that data is right-skewed.