Hao-3

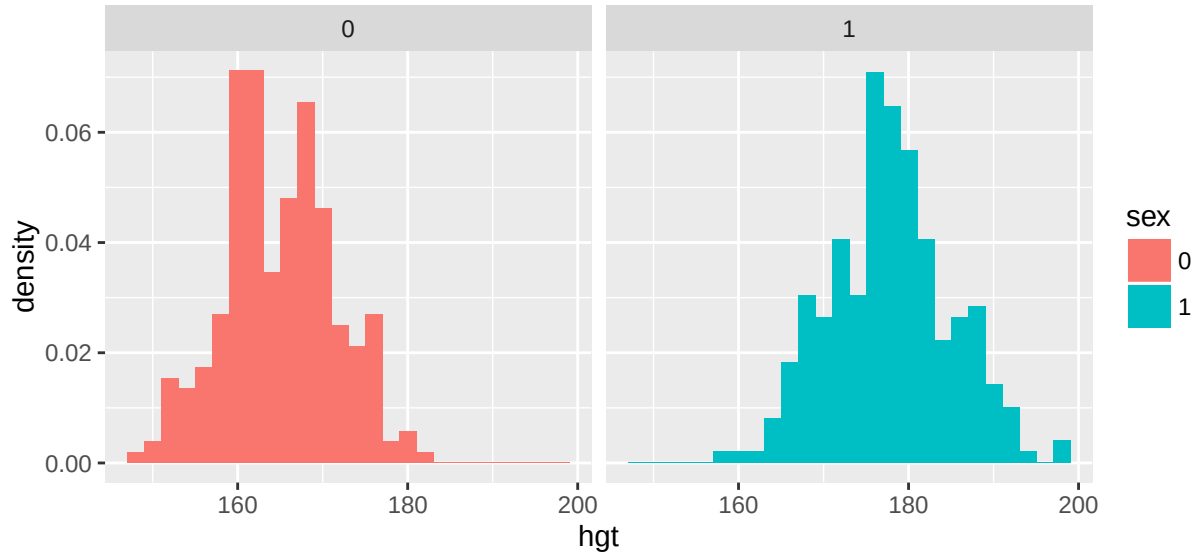
Bruce Hao

September 10, 2016

```
#setwd("~/Google Drive/CUNY/git/DATA606/Lab3")
#setwd("C:/Users/bhao/Google Drive/CUNY/git/DATA606/Lab3")
library(dplyr)
library(ggplot2)
download.file("http://www.openintro.org/stat/data/bdims.RData", destfile = "bdims.RData")
load("bdims.RData")
```

Exercise 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?

```
bdims %>% ggplot(aes(x=hgt, fill=sex)) + geom_histogram(aes(y=..density..), binwidth = 2) + facet_grid(
```

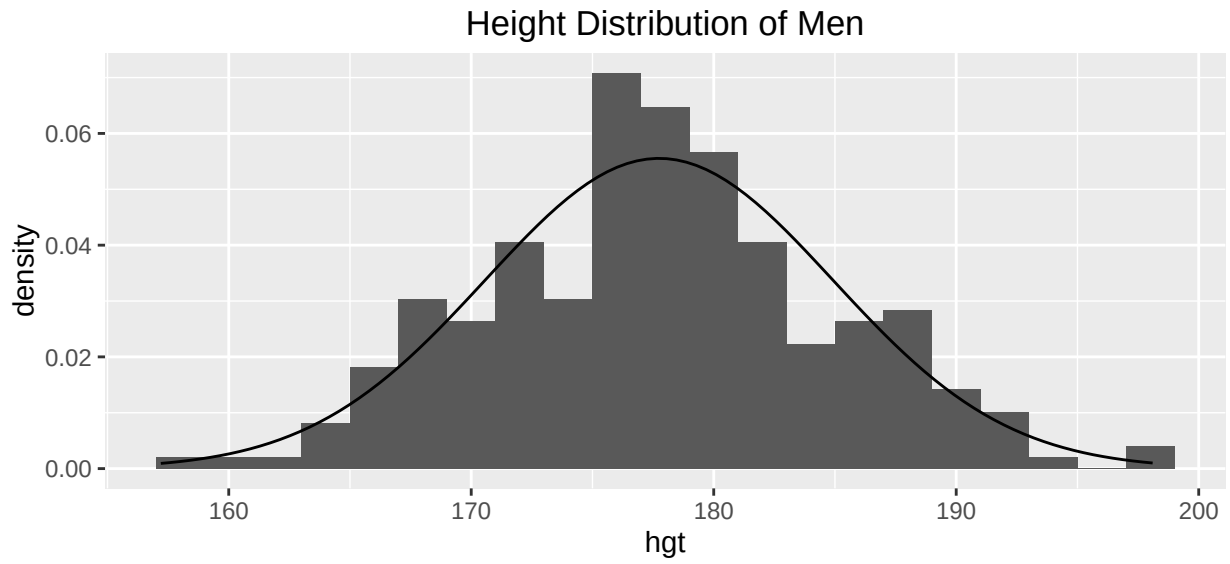


The distribution of men's heights is centered at a higher level and also more widely distributed than that of women's heights.

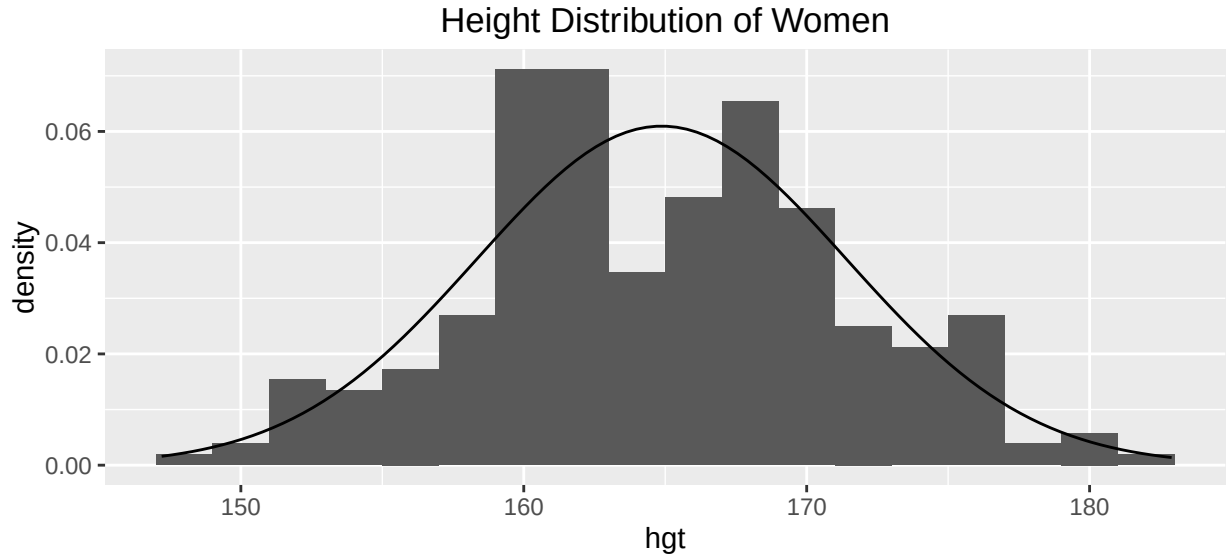
Exercise 2: Based on the this plot, does it appear that the data follow a nearly normal distribution?

Just eyeballing the histograms vs. the normal distribution, it appears that neither distribution is quite normal. The distribution of men's heights seems to have fatter tails (or at least a steeper center) relative to normal, and the distribution of women's heights seems to be bimodal.

```
# Men
bdims %>% filter(sex==1) %>% ggplot(aes(x=hgt)) + geom_histogram(aes(y=..density..), binwidth = 2) + sta
```



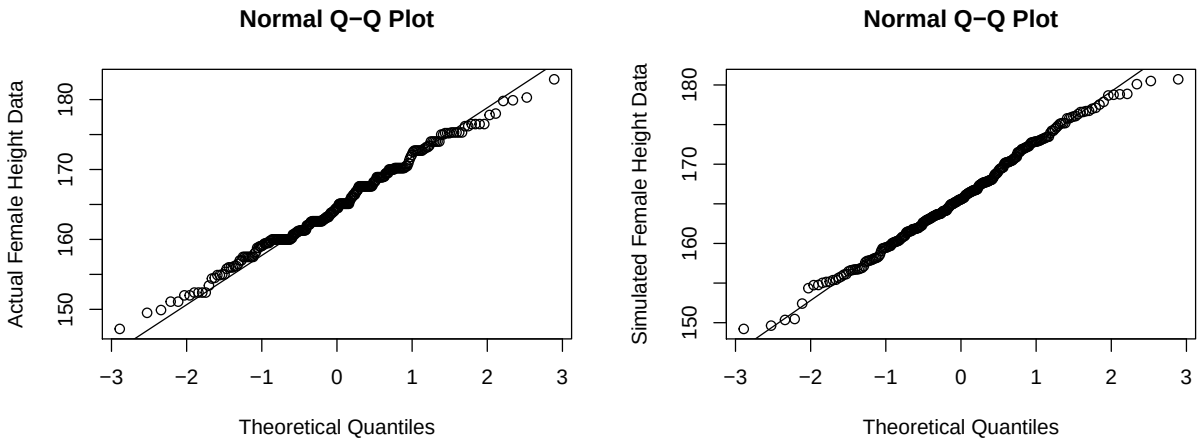

```r
# Women
fdims = bdims[bdims$sex==0, ]
bdims %>% filter(sex==0) %>% ggplot(aes(x=hgt)) + geom_histogram(aes(y=..density..), binwidth = 2) + sta
```



Exercise 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?

Even with the simulated data, not all of the points fall on the line. It's hard to tell if the simulated data fits teh line better than the actual data, which might suggest that the actual data is in fact normally distributed.

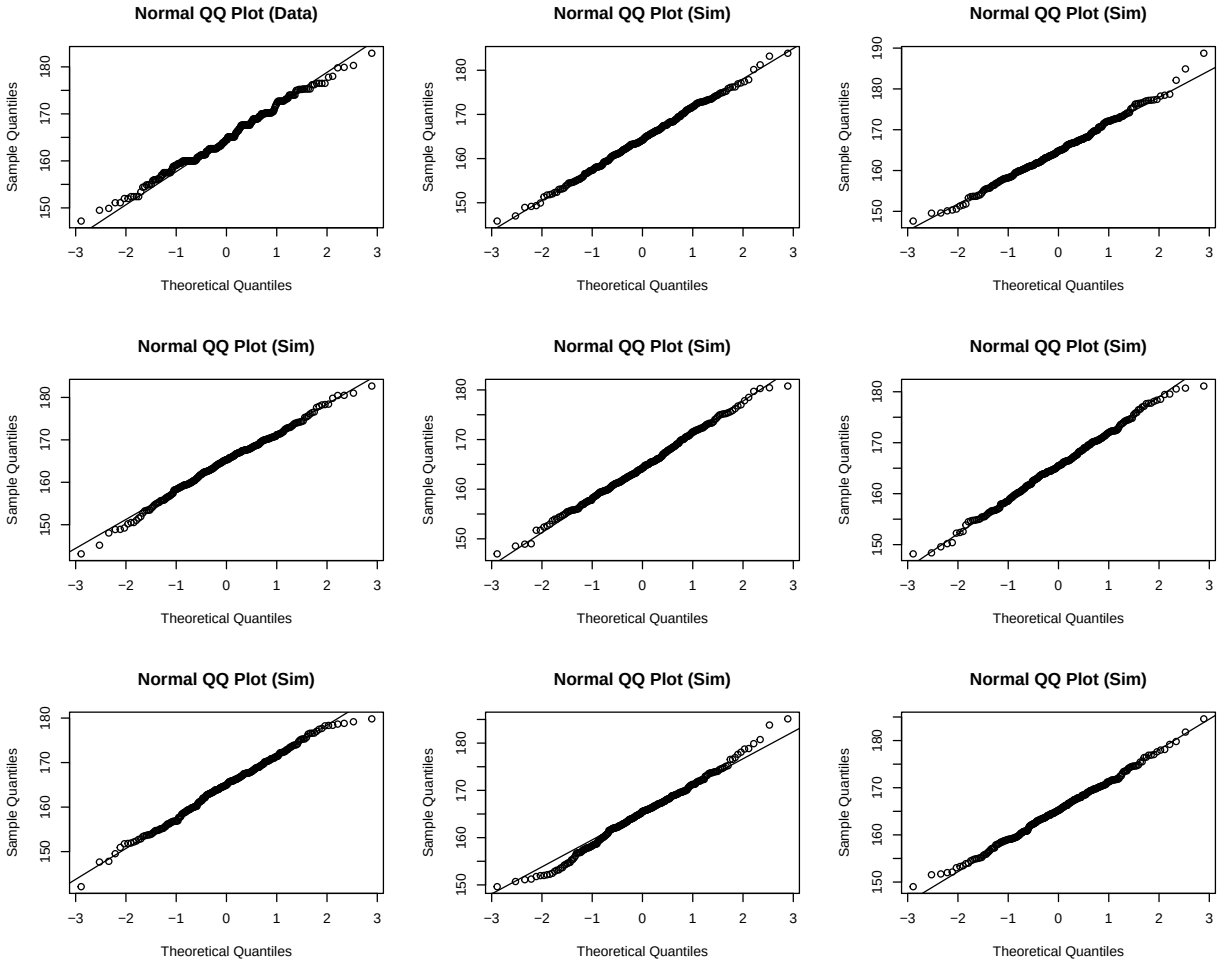
```r
sim_norm = rnorm(n = length(fdims$hgt), mean = mean(fdims$hgt), sd = sd(fdims$hgt))
par(mfrow=c(1,2))
qqnorm(fdims$hgt, ylab = 'Actual Female Height Data')
qqline(fdims$hgt)
qqnorm(sim_norm, ylab = 'Simulated Female Height Data')
qqline(sim_norm)
```



Exercise 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?

Aside from looking slightly more like a 'stair case' (probably given the resolution of the data), the actual data does very much resemble the simulated normally distributed data.

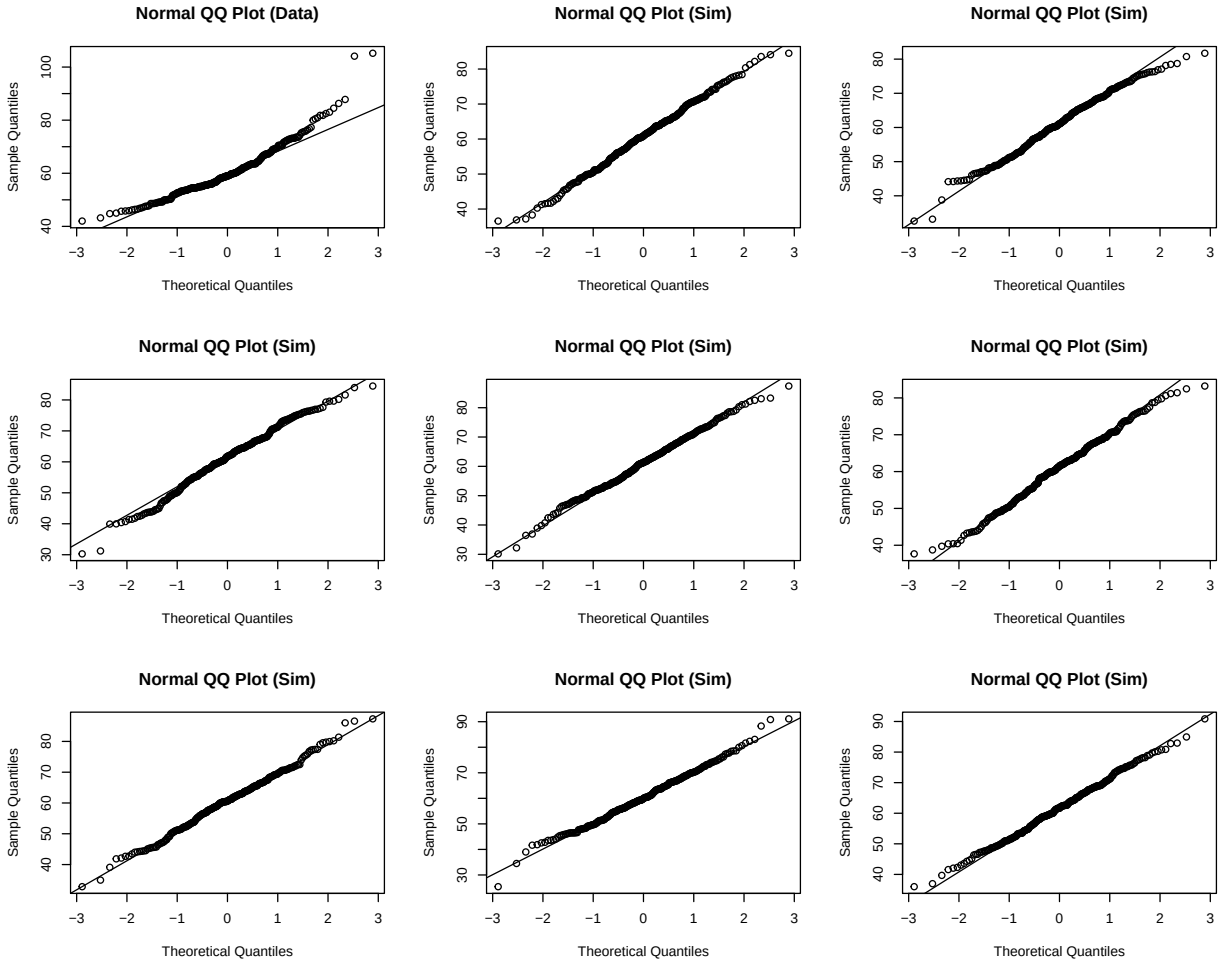
```
qqnormsim(fdims$hgt)
```



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

Again, the qq-plot for the actual female weight data does not appear significantly different from the qq-plots based on simulated data; however, there are a couple of outliers (104.1 and 105.2 kg observations) that fall much further from the line than do any of the simulated data points.

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



Exercise 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?

1) What is the probability that a woman's weight exceeds 80 kg? In both cases, the probability is approximately 2%.

```
sim_fwgt = rnorm(n = length(fdims$wgt), mean = mean(fdims$wgt), sd = sd(fdims$wgt))
1 - pnorm(q = 80, mean = mean(fdims$wgt), sd = sd(fdims$wgt))
```

```
## [1] 0.02182199
```

```
1 - pnorm(q = 80, mean = mean(sim_fwgt), sd = sd(sim_fwgt))
```

```
## [1] 0.0181619
```

2) What is the probability that a woman's height is below 155 cm? Again in both cases, the probabilities are about the same beetween 6-7%. According to the data set description, the measurements were taken from physically active individuals, mostly on the San Jose State and U.S. Naval Postgraduate

School campuses. It would have been interesting if the authors also collected information on ethnicity, as I would expect that a much higher percentage of women would be below 155 cm if the data were collected from certain other regions in the world.

```
sim_fhgt = rnorm(n = length(fdims$hgt), mean = mean(fdims$hgt), sd = sd(fdims$hgt))
pnorm(q = 155, mean = mean(fdims$hgt), sd = sd(fdims$hgt))
```

```
## [1] 0.06571769
```

```
pnorm(q = 155, mean = mean(sim_fhgt), sd = sd(sim_fhgt))
```

```
## [1] 0.05838601
```

On Your Own:

1a) Female biiliac diameter maps to qq-plot B; the left skew shows up on the qq-plot as the dots falling below the line in the bottom left quadrant.

1b) Female elbow diameter maps to qq-plot C.

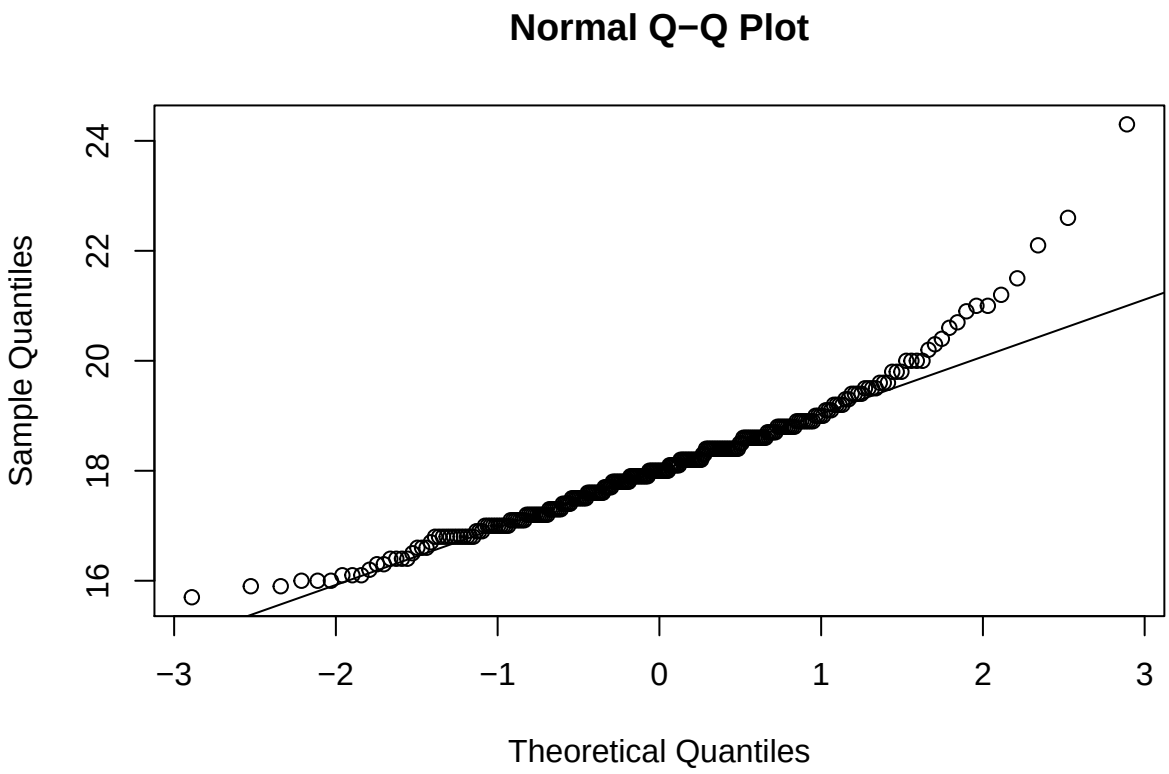
1c) Age maps to qq-plot D; the right skew shows up on the qq-plot as the dots falling above the line in the top right quadrant.

1d) Female chest depth maps to qq-plot A; the skew again shows up as the dots falling above the line in the top right quandrant, but the dots are otherwise much closer to the line suggesting a more normal distribution vs. that of age.

2) The slight step-wise pattern in qq-plots C and D likely have to do with the resolution at which the data was measured. For example, age was measured at the year level. As such, there are a lot of people with the exact same age vs. the number of people who have the exact same height at the 0.1 cm resolution level.

3) The normal plot of female knee diameter suggests that the data exhibit right skew.

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


Normal Q−Q Plot
Sample Quantiles
16
18
20
22
24
−3
−2
−1
0
1
2
3
Theoretical Quantiles