da460\_lab3\_grahn

jason grahn

10/8/2018

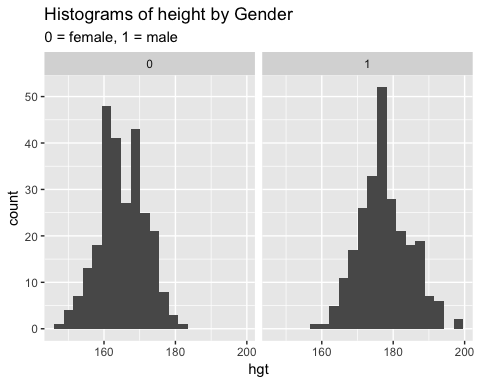
And we’ll also take the subsets from the lab handout, but make them using tidy verbs.

#filter the bdims data for sex == 1 because those are males.  
#mdims <- subset(bdims, sex == 1)  
mdims <- bdims %>%   
 filter(sex == 1)  
  
#filter the bdims data for sex == 0 because those are females  
#fdims <- subset(bdims, sex == 0)  
fdims <- bdims %>%   
 filter(sex == 0)

# 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?

#likely the instruction wants us to do this with each of the datasets, but we can do this with GGplot in one swoop.  
bdims %>%   
 ggplot(aes(x = hgt)) +   
 geom\_histogram(bins = 20) +   
 facet\_grid(. ~ sex) +  
 labs(title = "Histograms of height by Gender",  
 subtitle = "0 = female, 1 = male")

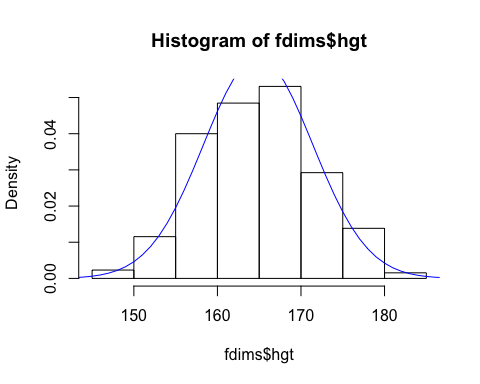


Both histograms seem to have a unimodal distribution, but the female histogram is a bit more smooth, though it does have a less “peaky” high-point. The histogram for men is distinctly shifted to higher numbers. Simply looking at the graphs, the spread appears similar, without any particular skewness to either.

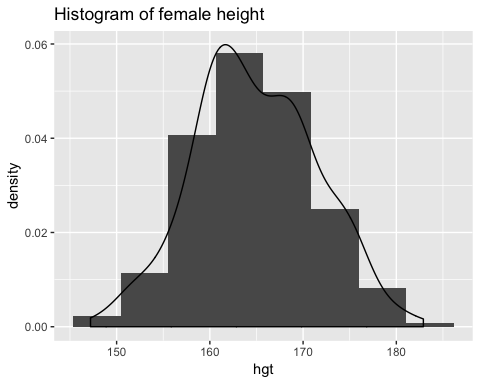
# Exercise 2

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

hist(fdims$hgt, probability = TRUE)  
x <- 140:190  
y <- dnorm(x = x, mean = fhgtmean, sd = fhgtsd)  
lines(x = x, y = y, col = "blue")



ggplot(fdims, aes(x=hgt, y=..density..)) +   
 geom\_histogram(position="identity", bins = 8) +   
 geom\_density() +  
 labs(title = "Histogram of female height")



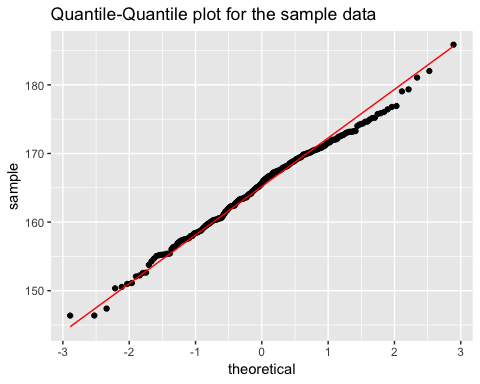
## Based on the this plot, does it appear that the data follow a nearly normal distribution?

Yes, based on the density histogram, these appear to nearly follow a normal distribution, though not perfect.

# Exercise 3

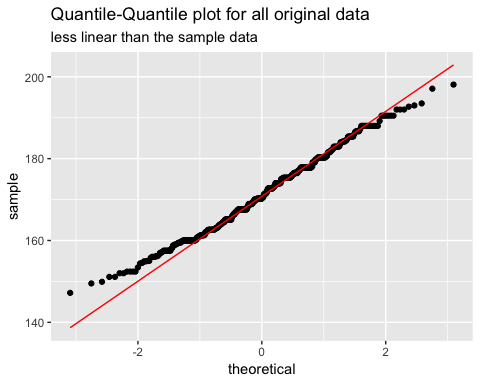
## Make a normal probability plot of sim\_norm. Do all of the points fall on the line? How does this plot compare to the probability plot for the real data?

#this is how we do normal probability plots in baseR.  
#qqnorm(fdims$hgt)  
#qqline(fdims$hgt)  
  
#simulating a normal distribution, per the lab  
sim\_norm <- rnorm(n = length(fdims$hgt), mean = fhgtmean, sd = fhgtsd)  
  
#and here's how we can do the QQplot in tidy; which allows us different kinds of flexibility based on the data  
tibble(sim\_norm) %>%   
 ggplot(aes(sample = sim\_norm)) +   
 stat\_qq() +  
 stat\_qq\_line(color = "red") +  
 labs(title = "Quantile-Quantile plot for the sample data")



No, not all points fall exactly on the line, but they aren’t terribly far off! To view the probability plot for the original data, we run the same format code…

#load the bdims data  
bdims %>%   
 #build your ggplot interface  
 ggplot(aes(sample = hgt)) +   
 #quantile-quantile layer  
 stat\_qq() +  
 #slope and intercept layer  
 stat\_qq\_line(color = "red") +  
 labs(title = "Quantile-Quantile plot for all original data",  
 subtitle = "less linear than the sample data")

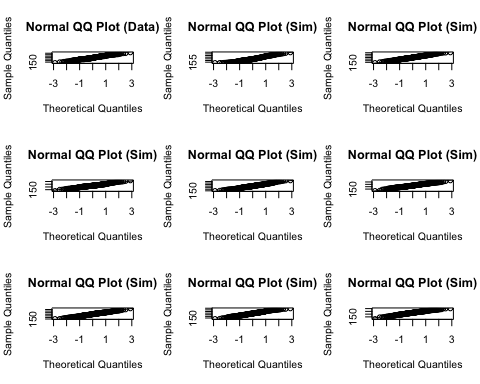


…and end up with a QQplot that, while still generally normal, shows definite variation differences than the sample data. The points end up on the line less - not *just* at the tails, but in the center areas as well.

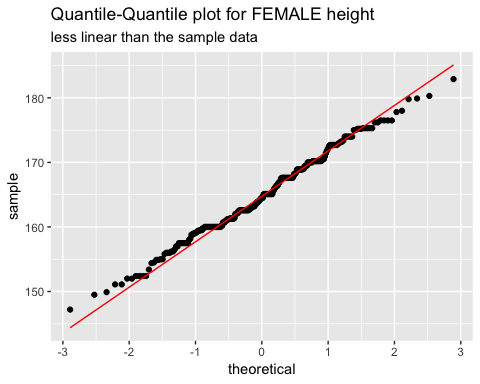
# Exercise 4

## 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?

#first we load the simulated data plots  
#qqnormsim is a custom function from OpenIntro  
qqnormsim(fdims$hgt)



fdims %>%   
 #build your ggplot interface  
 ggplot(aes(sample = hgt)) +   
 #quantile-quantile layer  
 stat\_qq() +  
 #slope and intercept layer  
 stat\_qq\_line(color = "red") +  
 labs(title = "Quantile-Quantile plot for FEMALE height",  
 subtitle = "less linear than the sample data")

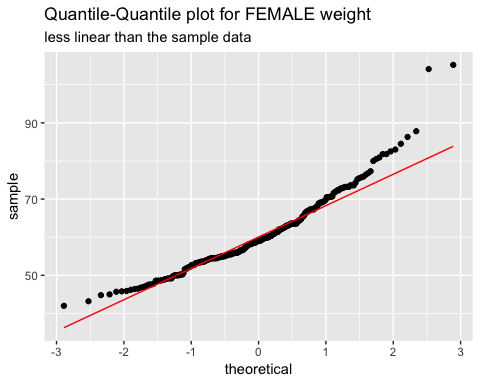


Yes, the normal probability plot for fdims$hgt looks similar to the plots created for the simulated data. They are both approximately linear with some variation at the tails. The plot for fdims$hgt supports that the female height distributino is approximately normal.

# Exercise 5

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

#qqplot   
fdims %>%   
 #build your ggplot interface  
 ggplot(aes(sample = wgt)) +   
 #quantile-quantile layer  
 stat\_qq() +  
 #slope and intercept layer  
 stat\_qq\_line(color = "red") +  
 labs(title = "Quantile-Quantile plot for FEMALE weight",  
 subtitle = "less linear than the sample data")



#density histogram

The two plots above highlight female weights do *not* appear to come from a normal distrubution. The QQPlot shows a defined curvature, non-linear behavior, with each tail peaking upward while the center lies beneath the line. This implies something other than normal distribution. On either end there appear to be systematic outliers as well. The density histogram and line provide an appearance of poisson distribution.

# 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?

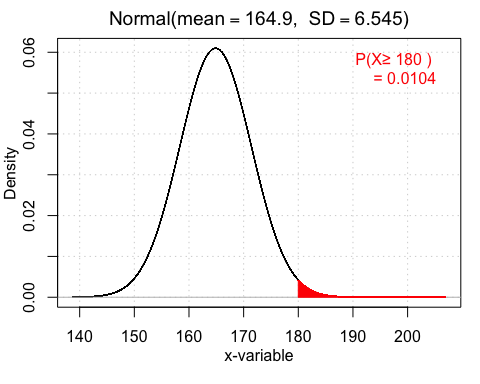
### Question 1: What is the probability a female height is greater than 180?

#### Theoretical normal

1 - pnorm(q = 180, mean = fhgtmean, sd = fhgtsd)

## [1] 0.01040328

iscamnormprob(180, mean = fhgtmean, sd = fhgtsd, direction = "above")



## probability: 0.0104

#### Empirical normal

sum(fdims$hgt > 180) / length(fdims$hgt)

## [1] 0.007692

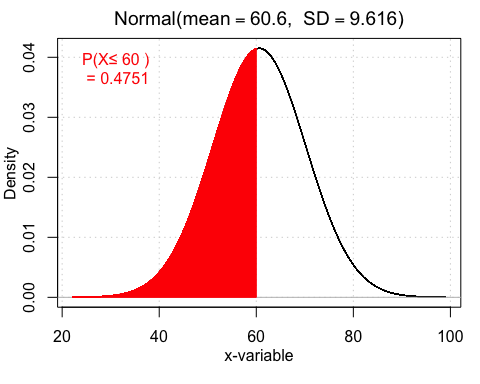
### Question 2: What is the probability a female weight is less than 60?

#### Theoretical normal

1 - pnorm(q = 60, mean = mean(fdims$wgt), sd = sd(fdims$wgt))

## [1] 0.5249

iscamnormprob(60, mean = mean(fdims$wgt), sd = sd(fdims$wgt), direction = "below")



## probability: 0.4751

#### Empirical normal

sum(fdims$wgt < 60) / length(fdims$wgt)

## [1] 0.5462

The height data was more accurate with an absolute difference of 0.0027; where as the weight data was further away at an absolute difference of 0.0213.