Week3- Lab III

Aamol Gote

9/18/2019

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

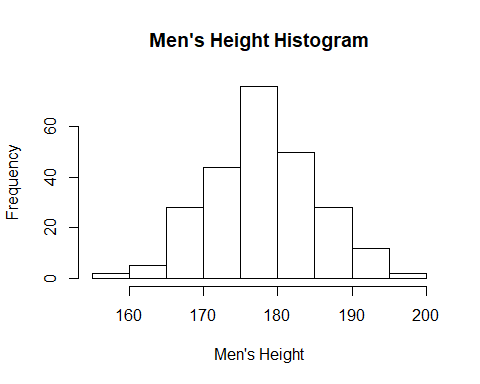
mdims <- subset(bdims, sex == 1)  
fdims <- subset(bdims, sex == 0)

***Excercise 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?

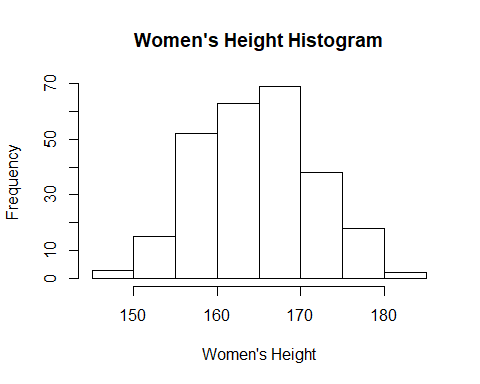
Men’s Height Histogram

hist(mdims$hgt,main="Men's Height Histogram", xlab = "Men's Height")



Women’s Height Histogramm

hist(fdims$hgt,main="Women's Height Histogram", xlab = "Women's Height")



mean(mdims$hgt)

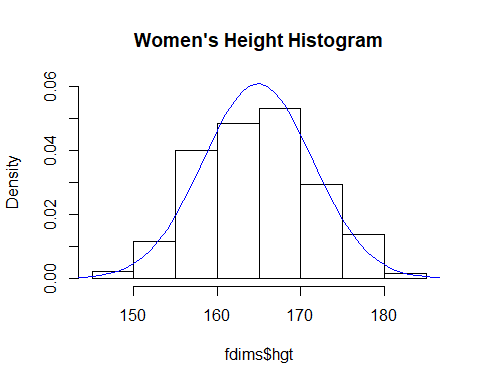
## [1] 177.7453

mean(fdims$hgt)

## [1] 164.8723

1. Men’s histogram is more accurate normal distribution bell curve than women’s.
2. Spread for women is less than men’s spread.
3. The mean of women is less than men which shifts the bell curve towards left. As means are different the bell curve centers are located differently.

fhgtmean <- mean(fdims$hgt)  
fhgtsd <- sd(fdims$hgt)  
hist(fdims$hgt, probability = TRUE, main = "Women's Height Histogram", ylim = c(0, 0.06))  
x <- 140:190  
y <- dnorm(x = x, mean = fhgtmean, sd = fhgtsd)  
lines(x = x, y = y, col = "blue")

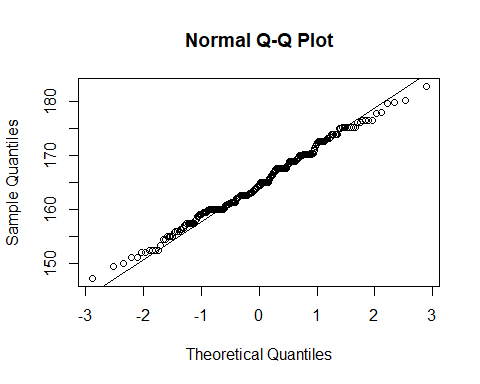


***Excercise 2***

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

It does follow normal distribution.

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

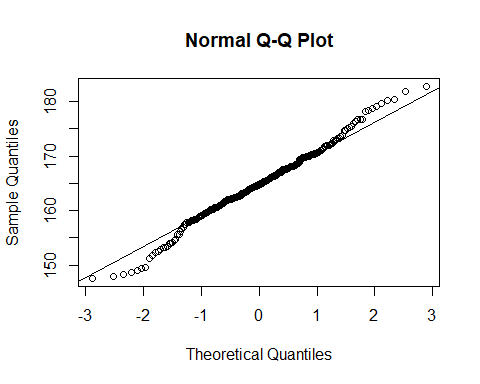


sim\_norm <- rnorm(n = length(fdims$hgt), mean = fhgtmean, sd = fhgtsd)

***Excercise 3***

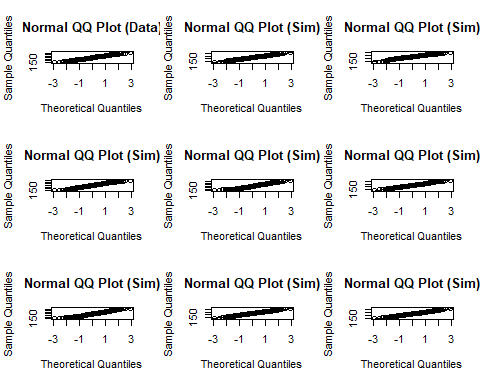
Make a normal probability plot of sim\_norm. Do all of the points fall on the line? How does this plot compare to the probability plot for the real data?

qqnorm(sim\_norm)  
qqline(sim\_norm)



1. Plot indicates that majority of the points tend to follow the line barring some at either tail ends.
2. Major difference comparing simulated data with actual data is that in simulated data tails are on the opposite side compared to that one with real data. Tials going in opposite direction is visible in standard deviation of 3 and -3.
3. Real data points have shape of of like stairs which is not evident in the simulated data, this is visible between standard deviation of 1 to -1 approximately.

qqnormsim(fdims$hgt)



***Excercise 4***

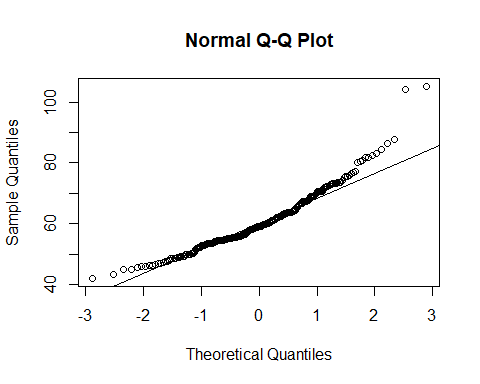
Does the normal probability plot for fdims$hgt look similar to the plots created for the simulated data? That is, do plots provide evidence that the female heights are nearly normal?

1. The qq plot for normal and simulated are smilar which proves that female heights are nearly normal.
2. One common theme across both simulated and real are the deviations of the points at the tails.
3. Stepped data points is evident in one of simulated data plot which exists in real data as well.

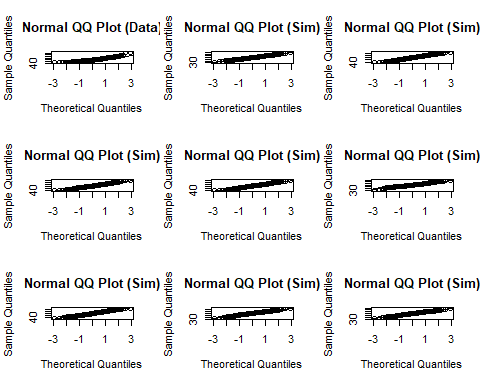
***Excercise 5***

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

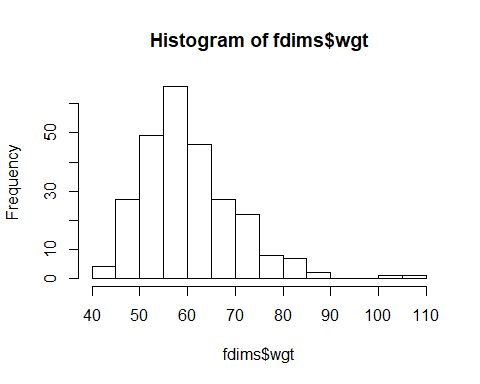
qqnorm(fdims$wgt)  
 qqline(fdims$wgt)



qqnormsim(fdims$wgt)



hist(fdims$wgt)



1. Female weights distribution follows normal distribution.
2. If you look at the actual data, majority of the points are near the line with exception of tails which is expected. Point distribution is slightly right skewed based on the QQ plot for actual data which can be evident from the histogram as well.

***Excercise 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. First Probability question - What’s the probablity that someone is less than 150 cm in height for females?
2. Theoretical Normal distribution

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

## [1] 0.01152955

1. Empirical distribution

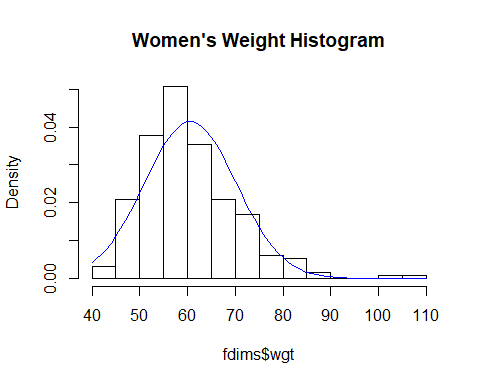
sum(fdims$hgt < 150) / length(fdims$hgt)

## [1] 0.01153846

Both theoretical and empirical distribution probability values are almost same.

Women’s Height Histogram

fwgtmean <- mean(fdims$wgt)  
 fwgtsd <- sd(fdims$wgt)  
 hist(fdims$wgt, probability = TRUE, main = "Women's Weight Histogram")  
 x <- 40:110  
 y <- dnorm(x = x, mean = fwgtmean, sd = fwgtsd)  
 lines(x = x, y = y, col = "blue")



1. Second Probability question - What’s the probablity that someone is greater than 70 KG in weight for females?
2. Theoretical Normal distribution

1 - pnorm(q = 70, mean = fwgtmean, sd = fwgtsd)

## [1] 0.1641539

1. Empirical distribution

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

## [1] 0.1576923

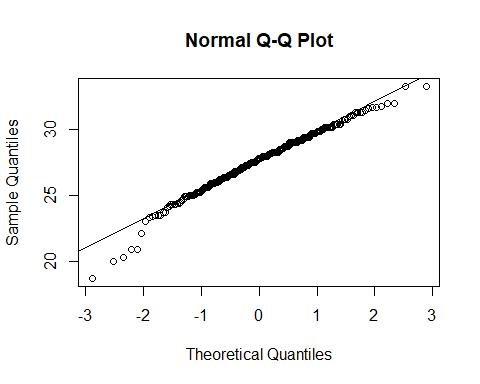
Both Values are closer but not closer compared to height.

1. Height number is in closer agreement than weight. Overall weight numbers for other probablities do not match much across 2 methods compared to height.

***On Your Own***

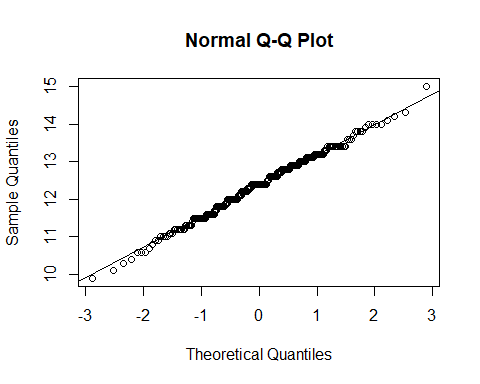
1. Now let’s consider some of the other variables in the body dimensions data set. Using the figures at the end of the exercises, match the histogram to its normal probability plot. All of the variables have been standardized (first subtract the mean, then divide by the standard deviation), so the units won’t be of any help. If you are uncertain based on these figures, generate the plots in R to check.
2. The histogram for female biiliac (pelvic) diameter (bii.di) belongs to normal probability plot letter **b**

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



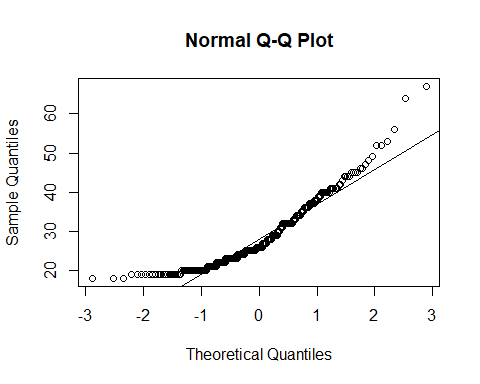
1. The histogram for female elbow diameter (elb.di) belongs to normal probability plot letter **c**

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



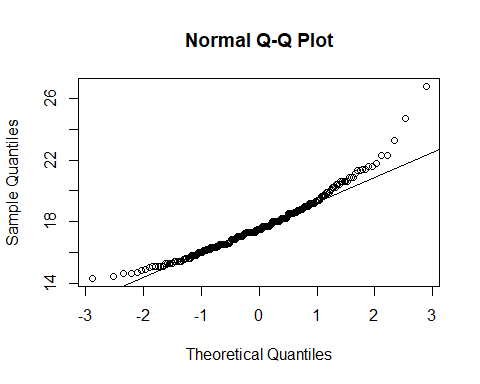
1. The histogram for general age (age) belongs to normal probability plot letter **d**

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



1. The histogram for female chest depth (che.de) belongs to normal probability plot letter **a**

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

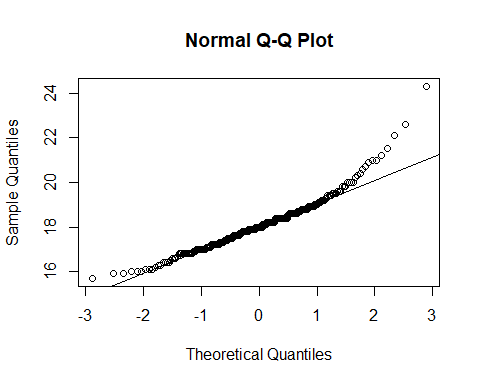


1. Note that normal probability plots C and D have a slight stepwise pattern.Why do you think this is the case?

Stepwise patterns is observed when data is discrete. Data type associated with C and D are different, elbow diameter is in decimals while age is whole number (integer).

1. 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). Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Use a histogram to confirm your findings.

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



This variable is **right skewed**, which is evident from the below histogram.

hist(fdims$kne.di)

