Lab3

Priya Shaji

2/18/2019

## The Normal Distribution

### The data

download the data

download.file("http://www.openintro.org/stat/data/bdims.RData", destfile = "bdims.RData")

load the data

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

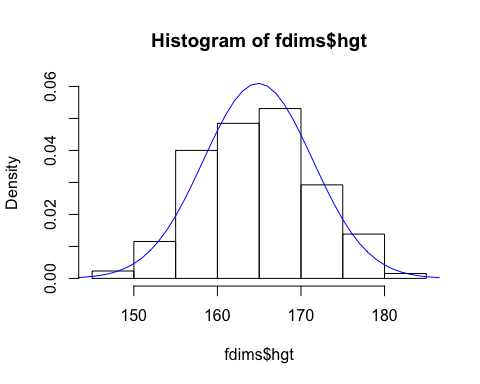
mdims <- subset(bdims, sex == 1)  
fdims <- subset(bdims, sex == 0)

The normal distribution analysis

creating density histograms

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

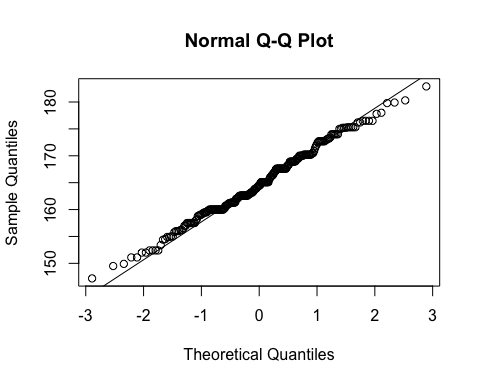
hist(fdims$hgt, probability = TRUE,ylim = c(0, 0.06))  
x <- 140:190  
y <- dnorm(x = x, mean = fhgtmean, sd = fhgtsd)  
lines(x = x, y = y, col = "blue")



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

Evaluating the normal distribution

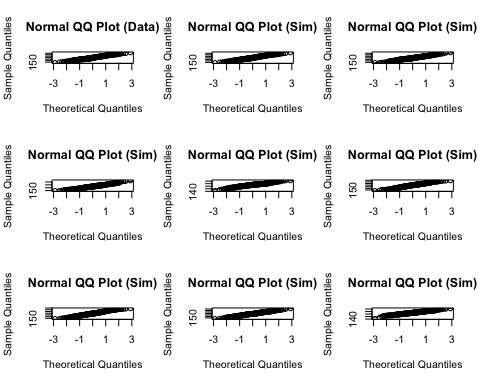
qqnorm(fdims$hgt)  
qqline(fdims$hgt)



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

## [1] 156.6163 177.5214 160.9691 164.4470 167.6804 165.0121 174.8099  
## [8] 169.5357 158.4031 163.3390 165.0829 166.1237 171.8333 171.9881  
## [15] 155.9922 165.7443 166.3368 163.0526 169.4194 169.0521 176.5394  
## [22] 164.0962 162.0378 165.6851 165.8738 165.2586 163.5606 169.8838  
## [29] 160.6067 157.2243 160.7929 160.6462 158.3445 165.4875 160.3583  
## [36] 166.1748 167.3081 170.4060 170.6082 161.3972 175.5315 167.8193  
## [43] 173.1248 157.0834 170.0288 171.4606 165.5835 168.6587 177.7264  
## [50] 160.7804 163.9177 164.8515 157.9473 167.3417 166.9616 151.9572  
## [57] 169.7347 174.0719 171.2858 172.0316 160.6335 170.6807 164.0375  
## [64] 172.2279 168.1213 159.7861 154.7593 160.2775 162.4151 171.8529  
## [71] 159.3997 173.7328 174.6528 156.8798 153.6818 170.9512 174.5223  
## [78] 153.6346 163.1282 178.8096 167.8733 159.7755 164.0437 154.7922  
## [85] 164.1118 171.1953 153.1964 156.5613 172.9219 158.9071 168.2042  
## [92] 159.1266 162.0083 166.9137 165.0597 160.3045 164.1252 163.7441  
## [99] 163.4541 152.2581 155.0797 166.1760 167.5998 154.4477 168.8341  
## [106] 181.0054 172.7982 154.8235 168.7062 164.8029 159.9024 173.6604  
## [113] 163.9666 174.5119 161.2008 186.6974 166.9362 145.3226 152.3457  
## [120] 168.1149 158.0330 165.0612 170.8900 154.3734 172.2707 172.0099  
## [127] 175.3428 167.6770 157.2926 163.6776 167.3131 159.7503 167.1142  
## [134] 169.1832 167.2509 165.6549 170.3005 157.4085 158.6583 161.3129  
## [141] 161.5258 161.1802 166.3513 168.5286 166.2617 156.5401 167.3539  
## [148] 163.0103 160.0850 165.8924 162.5971 166.7792 167.2066 168.1553  
## [155] 160.8537 161.4203 146.7916 173.9335 156.2121 160.7050 164.4111  
## [162] 154.4689 161.5353 162.4666 176.2301 154.0512 174.9430 161.9803  
## [169] 151.8757 171.7536 170.1319 165.2466 171.4465 170.9601 170.1360  
## [176] 174.3865 171.2489 166.5406 164.0898 158.8367 167.0728 159.9235  
## [183] 163.1113 164.0983 154.6402 147.0276 160.2315 159.4560 169.7723  
## [190] 171.4867 169.1853 164.3219 167.7898 166.6393 160.5060 155.1601  
## [197] 173.7050 170.5527 170.0083 164.4802 170.8854 154.6780 156.0484  
## [204] 177.7195 165.8035 181.3465 174.7404 147.1497 165.8883 153.2776  
## [211] 166.4764 165.2170 162.8460 163.2080 158.3460 170.1867 162.9215  
## [218] 163.4178 173.3326 165.9720 156.7970 162.5548 157.1881 154.5861  
## [225] 157.5739 166.9075 163.2508 169.6980 164.0974 166.4894 173.2401  
## [232] 173.2388 167.5596 168.8401 169.5764 173.1024 166.2004 169.3165  
## [239] 166.8919 168.5220 154.9998 173.2211 157.1935 171.2352 169.8465  
## [246] 164.9344 157.0533 164.8083 159.8916 165.9414 166.5885 175.2500  
## [253] 166.5891 159.3458 154.0118 173.8544 156.9925 167.4741 161.8056  
## [260] 167.4836

qqnormsim(fdims$hgt)



Normal probabilities

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

## [1] 0.004434387

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

## [1] 0.003846154

### Exercises

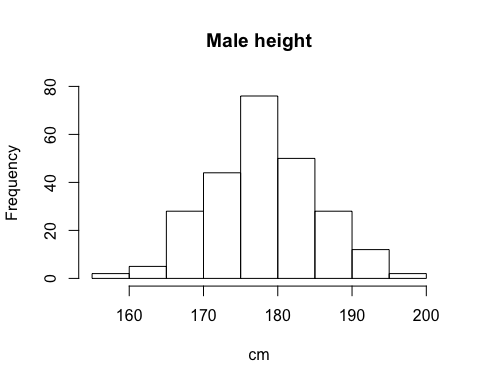
#### Exercise 1

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?

Answer 1

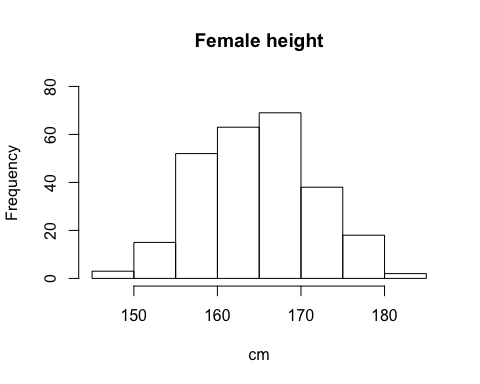
Men’s Height Histogram

hist(mdims$hgt, main="Male height", xlab="cm", ylim = c(0, 80))



Women’s Height Histogram

hist(fdims$hgt, main="Female height", xlab="cm", ylim = c(0, 80))



Most of the men’s height ranges between 175 cm to 180 cm with a fre quency of 80, where as most of women’s height range between 165cm to 170 cm with a frequency of 75.

#### Exercise 2

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

Answer 2

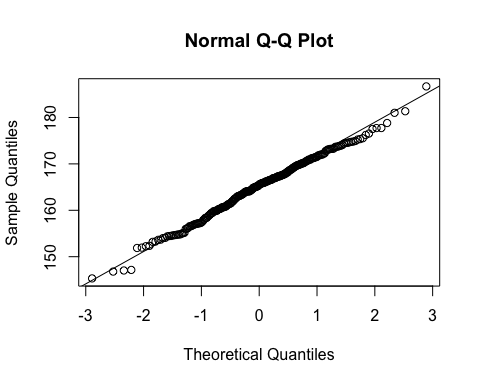
Yes, It looks like the data follows a nearly normal distribution

#### 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?

Answer 3

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



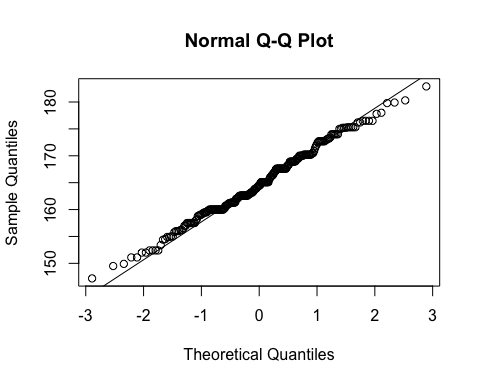
Yes, It seems like the data follows nearly normal distribution.The points are close to the line or the points are close to the real data.

#### 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?

Answer 4

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



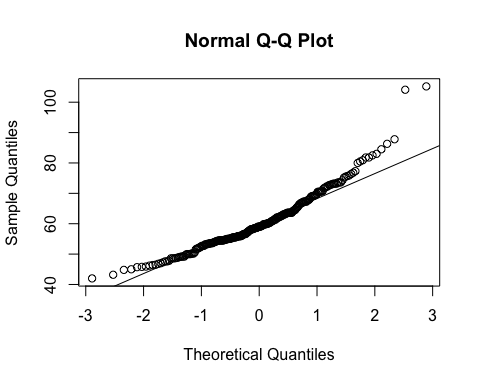
These plots are most likely similar. There are few variations along the y axis , though not much off as to think the distributions are not normal.

#### Exercise 5

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

Answer 5

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



The distribution for weight is right skewed . The Q-Q plot for the real data diverges from the line. Therefore the distribution does not seem normal.

#### 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?

Answer 6

Q: What is the probability that a random chosen young adult female is shorter than 170 cm?

pnorm(q = 170, mean = fhgtmean, sd = fhgtsd)

## [1] 0.7833331

sum(fdims$hgt < 170) / length(fdims$hgt)

## [1] 0.7538462

Q: Whats the probability that a female weighs less that 50kg?

fwgtmean <- mean(fdims$wgt)  
fwgtsd <- sd(fdims$wgt)  
pnorm(q=50,mean=fwgtmean,sd=fwgtsd)

## [1] 0.135143

sum(fdims$wgt < 50) / length(fdims$wgt)

## [1] 0.1038462

Generally, height is closer

### On Your Own

#### Ques\_1

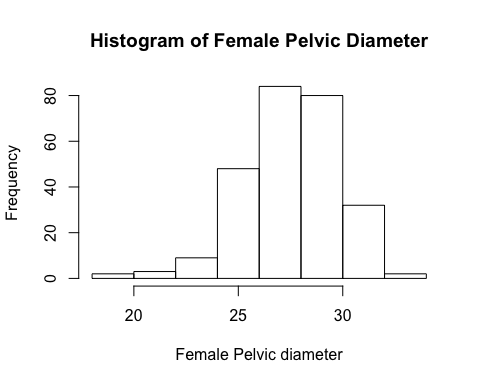
1: Now let’s consider some of the other variables in the body dimensions data set. Using the figures at the end of the exercises, match the histogram to its normal probability plot. All of the variables have been standardized (first subtract the mean, then divide by the standard deviation), so the units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

Answer 1

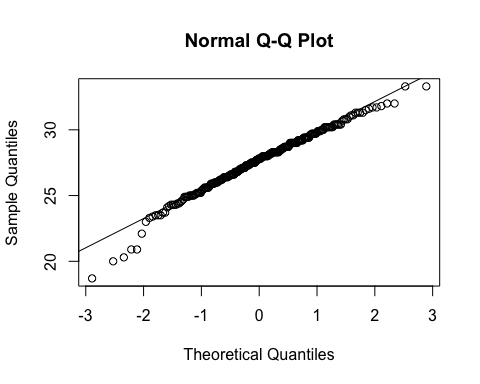
1. The histogram for female biiliac (pelvic) diameter (“bii.di”) belongs to normal probability plot letter:

PLOT B

hist(fdims$bii.di,xlab = "Female Pelvic diameter", main = "Histogram of Female Pelvic Diameter")



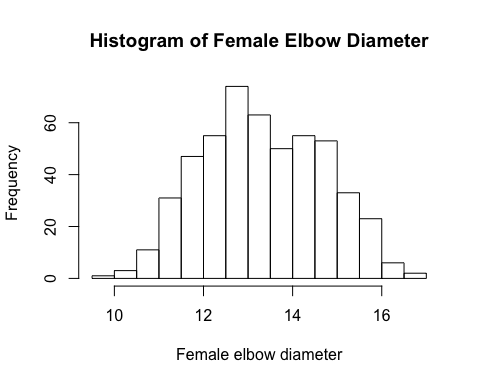
qqnorm(fdims$bii.di)  
qqline(fdims$bii.di)



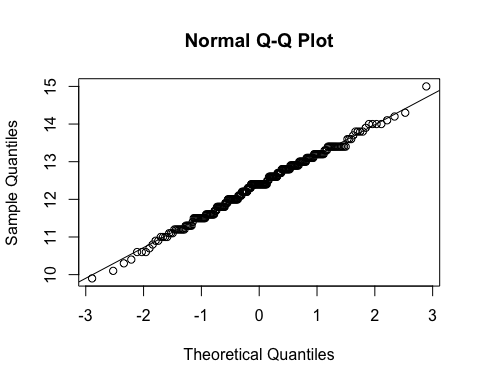
1. The histogram for female elbow diameter (elb.di) belongs to normal probability plot letter \_\_\_\_.

PLOT C

hist(bdims$elb.di,xlab = "Female elbow diameter", main = "Histogram of Female Elbow Diameter")



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



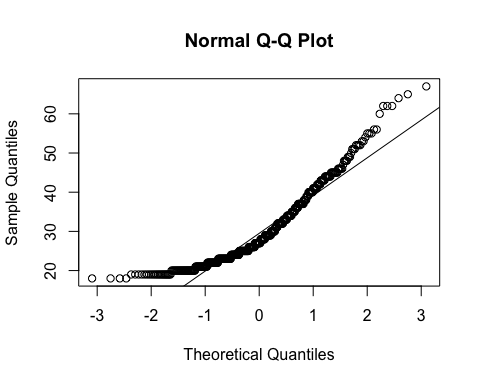
1. The histogram for general age (age) belongs to normal probability plot letter \_\_\_\_.

PLOT D

hist(bdims$age, xlab = "Age in years", main = "Histogram of Sample Age in Years")



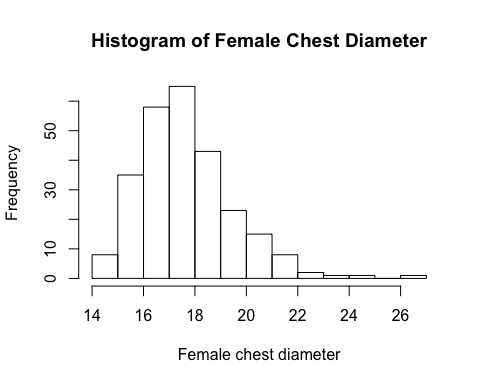
qqnorm(bdims$age)  
qqline(bdims$age)



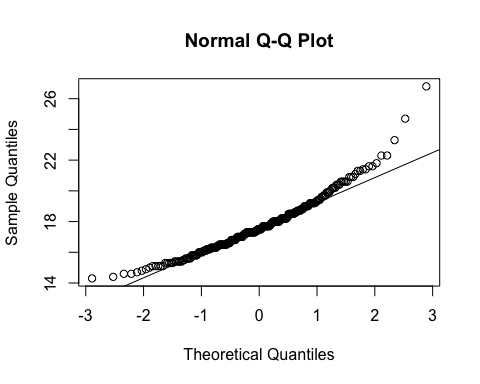
1. The histogram for female chest depth (che.de) belongs to normal probability plot letter \_\_\_\_.

PLOT A

hist(fdims$che.de,xlab = "Female chest diameter", main = "Histogram of Female Chest Diameter")



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



#### Ques\_2

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

Answer 2

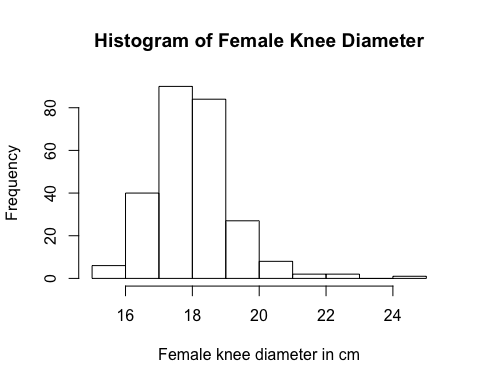
Likely because of the integer values provided in the data set. Age was given in integers making the jumps a bit more obvious. Perhaps for elbow diameters, many people cluster around the same diameter, since it functions in the same way for many people.

#### Ques\_3

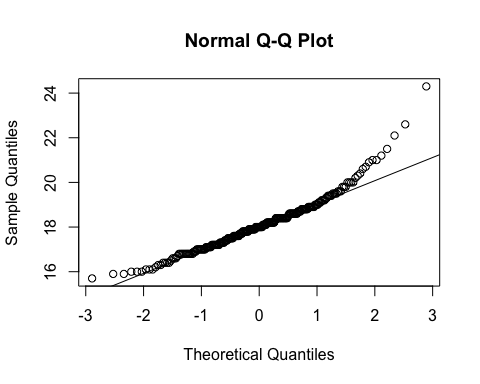
1. As you can see, normal probability plots can be used both to assess normality and visualize skewness. Make a normal probability plot for female knee diameter (kne.di). Based on this normal probability plot, is this variable left skewed, symmetric, or right skewed? Use a histogram to confirm your findings.

Answer 3

hist(fdims$kne.di, xlab = "Female knee diameter in cm", main = "Histogram of Female Knee Diameter")



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



From the probability plot it appears there are less values as the quantites increase. Therefore, it appears as through it is skewed right.