Week 1 HW Submission

Paul Isaacson

January 23, 2017

## Exercise 1

For this exercise, you'll explore and summarize data on cholesterol levels for 40 randomly selected American women.

### Part 1a

From the HealthExam data set, extract the cholesterol level of the 40 women and assign it to the variable fs. As a safety check, the mean cholesterol level should be 240.875 mg/dl if you've done it correctly.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

fs <- HealthExam$Cholesterol[HealthExam$Sex == 'F']

### Part 1b

Apply summary() and sd() to the vector fs to find summary statistics for female cholesterol level. Based on the summary statistics, does the data appeared to be skewed? If so, which direction?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1b -|-|-|-|-|-|-|-|-|-|-|-

summary(fs)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.0 120.2 194.0 240.9 303.0 920.0

sd(fs)

## [1] 185.9824

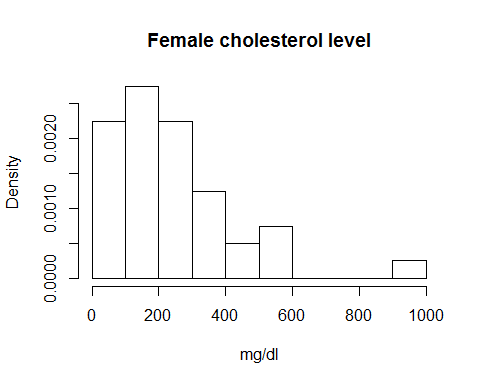
Right Skewed. The max is large at 920.0. Also, the mean is higher than the median, another indicator of skew.

### Part 1c

Make a histogram for female cholesterol level. Label the -axis with "mg/dl" and title the plot "Female cholesterol level". Does the shape of the distribution agree with your answer to Exercise 3? Based on the histogram, does the variable female cholesterol level appear to be approximately normally distributed? Explain.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

hist(fs, xlab='mg/dl', main='Female cholesterol level', prob=TRUE)



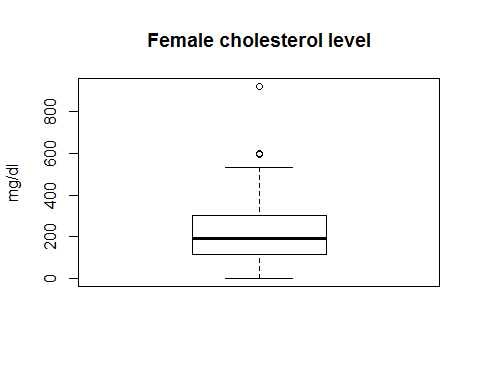
The shape also indicates right-skew. The data does not appear to fit a normal distribution. The mound shape is missing the left side.

### Part 1d

Make a boxplot for the same data as in Exercise 4. Label the -axis with "mg/dl" and title it as before.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1d -|-|-|-|-|-|-|-|-|-|-|-

boxplot(fs, ylab='mg/dl', main='Female cholesterol level')



### Part 1e

According to the 1.5 IQR rule, what is the largest value of female cholesterol level that would not be considered an outlier?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e -|-|-|-|-|-|-|-|-|-|-|-

quantile(fs, .75) + 1.5 \* IQR(fs)

## 75%   
## 577.125

### Part 1f

The maximum female cholesterol level is an outlier, find its -score. You'll need to combine the commands max(), mean(), and sd(). If the data did come from a normal distribution, would the maximum female cholesterol level seem unusually large? Explain.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1f -|-|-|-|-|-|-|-|-|-|-|-

(max(fs) - mean(fs)) / sd(fs)

## [1] 3.651556

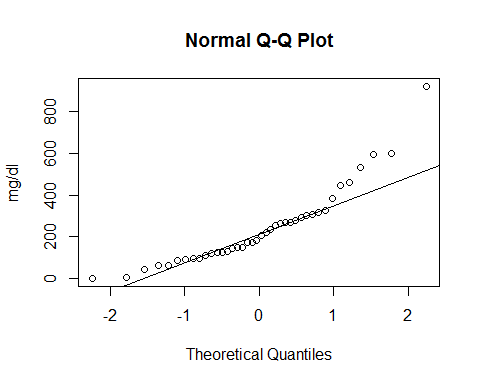
This is unusually large. The value of 3.651556 has a probability of 2.606562510^{-4} which in a population of 40 we would only expect to see 0.0104262 that are at least that many standard deviations away from the mean.

### Part 1g

Make a normal probability plot for fs, add a fit line to it, and label -axis appropriately. Based on the information in the plot, does this sample data seem to come from a normal distribution? Explain.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1g -|-|-|-|-|-|-|-|-|-|-|-

qqnorm(fs, ylab='mg/dl')  
qqline(fs)



No. On the lower end, the points are above the line indicating the data is the the right of where it would be in a normal distribution. Towards the upper end, the points are significantly higher than would be expected indicating a longer than expected right side.

## Exercise 2

This is essentially problem 3.11 from Chapter 3 in our textbook by Ott. We want to compare home ownership percentages for three different years: 1985, 1996, 2002.

### Part 2a

The code below loads a data set with randomly sampled test scores from three different instructors. Modify the code to load and analyze the home ownership percentages in the "homes.csv" file and use the plots to answer the questions below. Ott says to make relative frequency histograms (divide the frequencies by the sample size to get proportions), but we'll use density histograms instead (rescale the vertical so that the total geometric area in the bars is exactly 1) ... this makes it possible to compare histograms using different sample sizes and possibly different bins or classes.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2a -|-|-|-|-|-|-|-|-|-|-|-

homes <- read.csv("homes.csv")  
head(homes)

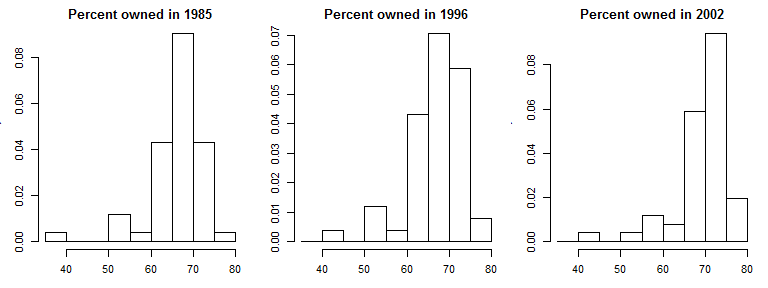
## PercentOwned Year  
## 1 70.4 year1985  
## 2 61.2 year1985  
## 3 64.7 year1985  
## 4 66.6 year1985  
## 5 54.2 year1985  
## 6 63.6 year1985

#summary(homes$PercentOwned[homes$Year=="year1985"])  
#summary(homes$PercentOwned[homes$Year=="year1996"])  
#summary(homes$PercentOwned[homes$Year=="year2002"])  
  
tapply(homes$PercentOwned, homes$Year, summary)

## $year1985  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 37.40 63.15 67.90 65.88 69.95 75.90   
##   
## $year1996  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 40.40 64.55 68.20 66.84 71.20 76.50   
##   
## $year2002  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 44.10 67.25 70.20 69.45 73.50 77.30

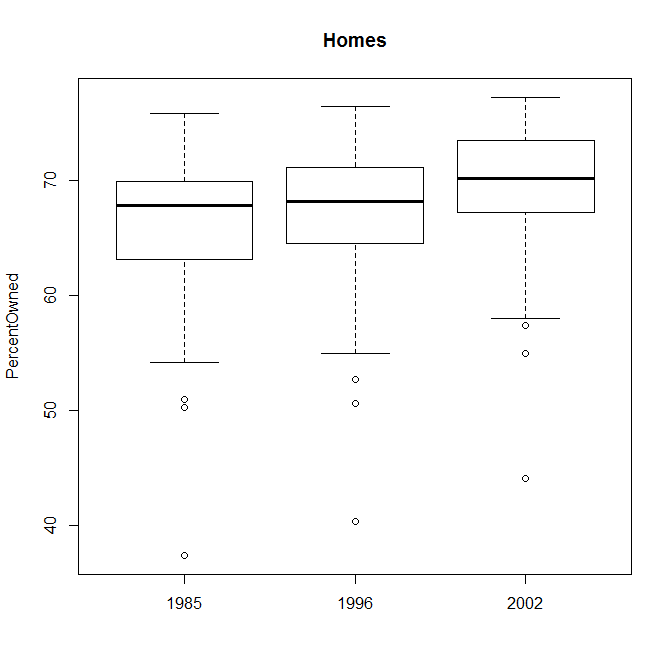
bins <- seq(35, 80, by=5)

par(mfrow=c(1,3),mar=c(3,3,2,1))  
  
hist(homes$PercentOwned[homes$Year=="year1985"],main="Percent owned in 1985",breaks = bins,xlab="PercentOwned", prob=TRUE)  
hist(homes$PercentOwned[homes$Year=="year1996"],main="Percent owned in 1996",breaks = bins,xlab="PercentOwned", prob=TRUE,ylab="")  
hist(homes$PercentOwned[homes$Year=="year2002"],main="Percent owned in 2002",breaks = bins,xlab="PercentOwned", prob=TRUE)



par(mfrow=c(1,1))

boxplot(homes$PercentOwned~homes$Year,names=c("1985","1996","2002"),ylab="PercentOwned",main="Homes")



### Part 2b

Comment on similarities and differences among the distributions of home ownership percentages for the years 1985, 1996, and 2002. Is there a general trend?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2b -|-|-|-|-|-|-|-|-|-|-|-

All three distributions are skewed to the left. The overall trend is that homeownership is going up.

## Exercise 3

Assume that the length of a natural human pregnancy is approximately normally distributed with mean 268 days and standard deviation 16 days. Use R to answer some questions about this distribution:

### Part 3a

Find the probability that a random natural pregnancy lasts less than 250 days.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3a -|-|-|-|-|-|-|-|-|-|-|-

pnorm(250, 286, 16)

## [1] 0.01222447

### Part 3b

Compute the probability that a natural human pregnancy lasts between 260 and 280 days.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3b -|-|-|-|-|-|-|-|-|-|-|-

pnorm(280, 286, 16) - pnorm(260, 268, 16)

## [1] 0.04529269

### Part 3c

How long are the longest 10% of natural human pregnancies?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3c -|-|-|-|-|-|-|-|-|-|-|-

qnorm(.10, 286, 16)

## [1] 265.4952

## Exercise 4

This problem is to demonstrate how to use the normal probability plot to help judge the fit of a normal distribution.

### Part 4a

The file bodyFat.csv is included with the weekly download. Use read.csv(...) to read the file into a dataframe. This is an artificial data set meant to be bodyfat percentages of 250 random men. Show how to load the data and assign the bodyfat percentages to a vector called bfat.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4a -|-|-|-|-|-|-|-|-|-|-|-

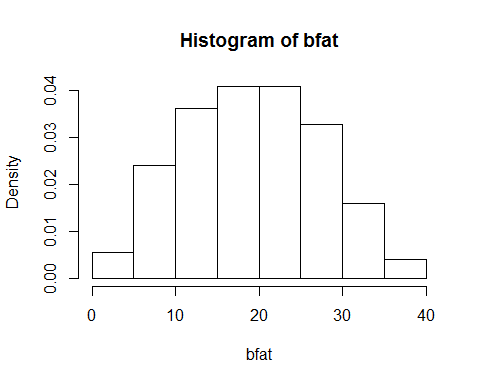
df <- read.csv('bodyFat.csv')  
bfat <- df$bodyFat

### Part 4b

Make a histogram of the bodyfat percentages. Does it appear that the data comes from a normally distributed random variable? Explain.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4b -|-|-|-|-|-|-|-|-|-|-|-

hist(bfat, prob=TRUE)



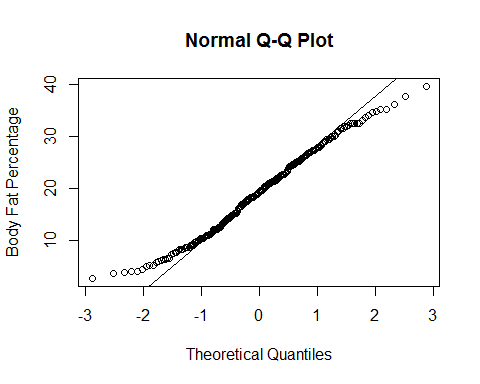
Yes, the data appears to come from a normal distribution.

### Part 4c

Now make a normal probability plot with a fitted line by using qqnorm() and qqline(). Note the "S" shape of the points. What is this telling you about the distribution of the bodyfat data?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4c -|-|-|-|-|-|-|-|-|-|-|-

qqnorm(bfat, ylab = 'Body Fat Percentage')  
qqline(bfat)



The s shape tells us that the left side is shifted to the right and the right side is shifted to the left. In short, the distribution is normal shape but more "squished"