Week 1 HW Submission

Matt Allen

5/30/2018

## How to complete the homework.

If you’ve made it this far, then you’ve already downloaded and unzipped the HW packet for this week. We suggest that you keep all of the materials, including this .rmd file, for the week in one folder. It will help to set the working directory to the folder that contains the HW materials. You can do this by opening the rmd file in an RStudio editor window and then using the menu commands Session -> Set Working Directory -> To Source File Location.

You’ll be adding R code and typing answers in the designated spaces throughout this document. At the end of the week you’ll submit the .rmd file and the “knitted” Word document to the dropbox on D2L.

Reminder:

## Exercise 1

For this exercise, you’ll explore and summarize data on cholesterol levels for 40 randomly selected American women. The dataset for this problem in the DS705data package. The code on lines 10-17 of this file makes sure that the package is installed when this file is knitted. The data() command at line 19 loads the dataset.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Assign cholestrol of 40 women in data set to variable fs.  
health\_exam\_female <- subset(HealthExam,HealthExam$Sex=="F")  
fs <- health\_exam\_female$Cholesterol  
# Check that the mean cholestrol level is 240.875  
mean(fs) == 240.875

## [1] TRUE

### 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 and sd on fs  
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

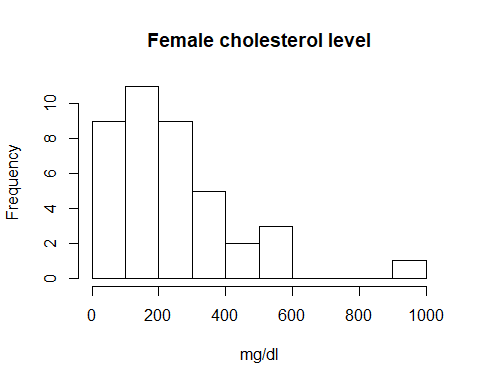
By inspection of the summary statisitics, the Mean appears significantly higher than the Median. This implies right skewness.

### 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 1b? Based on the histogram, does the variable female cholesterol level appear to be approximately normally distributed? Explain.

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

# histogram of female cholesterol level  
hist(fs,xlab="mg/dl",main = "Female cholesterol level")



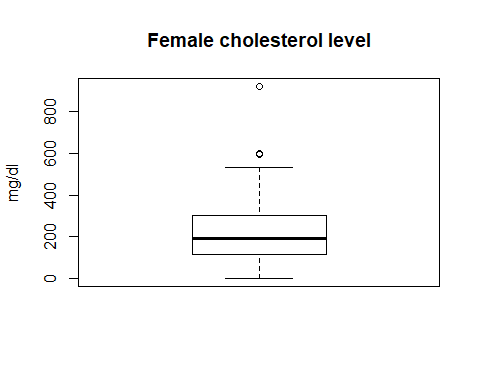
Based on the histogram, there is further evidence of right skewness. From the shape of the graph, it does not appear female cholesterol level is normally distributed.

### Part 1d

Make a boxplot for female cholesterol level. Label the -axis with “mg/dl” and title it as before.

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

# boxplot of female cholesterol level  
boxplot(fs,ylab="mg/dl",main = "Female cholesterol level")



### Part 1e

According to the 1.5 IQR rule, what is the cutoff value for outliers in the upper tail of female cholesterol level?

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

# quantile(fs,.75) gives the third quartile  
third\_quartile <- quantile(fs,.75)   
# IQR(fs) gives the interquartile range  
interquartile\_range <- IQR(fs)  
cutoff\_value <- third\_quartile + (1.5\*interquartile\_range)   
cutoff\_value

## 75%   
## 577.125

## From the 1.5 IQR Rule, the cutoff value is 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 -|-|-|-|-|-|-|-|-|-|-|-

# Find the z-score of female cholesterol level  
(max(fs) - mean(fs))/sd(fs)

## [1] 3.651556

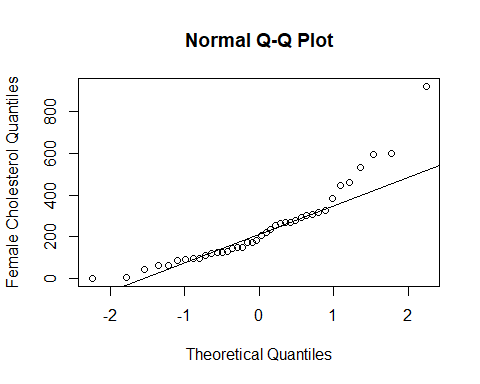
The z-score of 3.651556 of the maximum female cholesterol level would be unusually large if female cholesterol level was normally distributed. The z-score corresponds to standard deviations from 0 in the standard normal distribution. Anything beyond 3 standard deviations has a low probability of occurring.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# make a normal probability plot for fs and fit a line to it.  
{  
 qqnorm(fs, ylab="Female Cholesterol Quantiles")   
 qqline(fs)  
}



The normal probability plot is further confirmation that it is unlikely that female cholesterol is normally distributed. If it were normally distributed the dots would follow the line. It looks good in the middle, but moves away especially on the right.

## 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 (add the option prob=TRUE to the appropriate R command) … this makes it possible to compare histograms using different sample sizes and possibly different bins or classes.

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

# start by copying and pasting this block of code so you can refer to it as you tweak your own code  
# when you no longer want to execute this block you can delete it, or change the {r}  
# to {r, include=FALSE, eval=FALSE}. You can also minimize the whole block by clicking the little down arrow next to the line number at the beginning of the block.  
  
# load the data and look at it  
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

# summarize the PercentOwned for each Year  
summary(homes$PercentOwned[homes$Year=="year1985"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 37.40 63.15 67.90 65.88 69.95 75.90

sd(homes$PercentOwned[homes$Year=="year1985"])

## [1] 6.734407

summary(homes$PercentOwned[homes$Year=="year1996"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 40.40 64.55 68.20 66.84 71.20 76.50

sd(homes$PercentOwned[homes$Year=="year1996"])

## [1] 6.688027

summary(homes$PercentOwned[homes$Year=="year2002"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 44.10 67.25 70.20 69.45 73.50 77.30

sd(homes$PercentOwned[homes$Year=="year2002"])

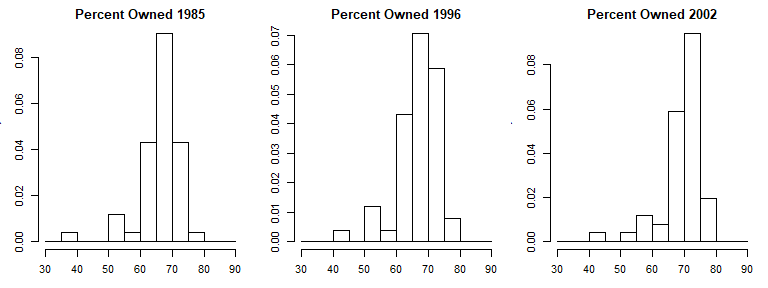
## [1] 6.162901

# a more advanced version of this summary by Year would be to use tapply() to loop over the factors.  
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

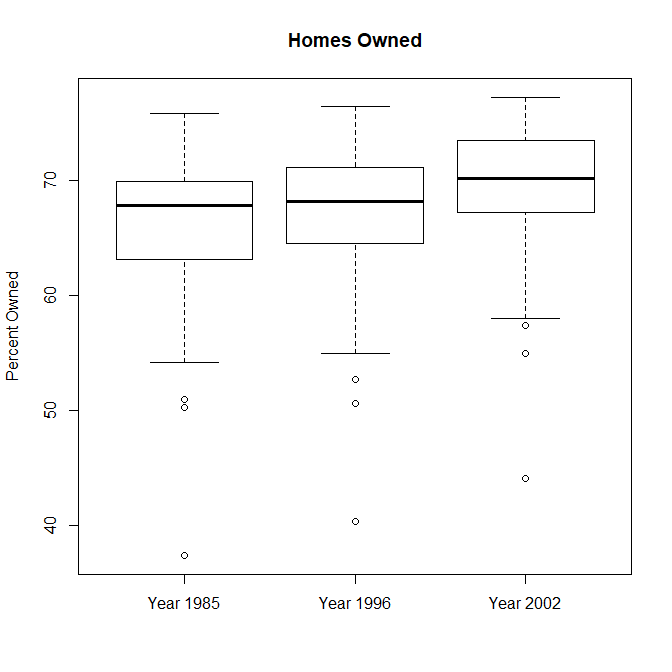
# looks like the PercentOwned across all three years range from 37.4 to 77.3, to make it easier to compare the distributions lets use the same bins for all three histograms. I chose a range from 30 to 90, because it left space on the sides of the histograms.  
bins <- seq(30,90,by=5)

# set up R to expect an array of plots with 1 row and 3 columns  
# the mar parameter adjust white space around the plot, notice that it has covered the axis labels which is OK here  
par(mfrow=c(1,3),mar=c(3,3,2,1))  
hist(homes$PercentOwned[homes$Year=="year1985"],main="Percent Owned 1985",breaks = bins,xlab="Percent Owned", prob=TRUE)  
hist(homes$PercentOwned[homes$Year=="year1996"],main="Percent Owned 1996",breaks = bins,xlab="Percent Owned", prob=TRUE,ylab="")  
hist(homes$PercentOwned[homes$Year=="year2002"],main="Percent Owned 2002",breaks = bins,xlab="Percent Owned", prob=TRUE)



# reset to one plot  
par(mfrow=c(1,1))

# make side-by-side boxplots to make comparisons easier  
boxplot(homes$PercentOwned~homes$Year,names=c("Year 1985","Year 1996","Year 2002"),ylab="Percent Owned",main="Homes Owned")



### 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 -|-|-|-|-|-|-|-|-|-|-|-

The shapes of the distributes look similar. It is possible that they are all approximately normally distributed, but they all have points that are small enough to have a low probability of occurring if normally distributed. The distributions may be somewhat left skewed. There is a general trend that mean and median percentage of home ownership is increasing over time.

## 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 -|-|-|-|-|-|-|-|-|-|-|-

# probability that natural pregnancy last less that 250 days  
pnorm(250, mean = 268, sd = 16)

## [1] 0.1302945

## There is a 13% chance of having a natural pregnancy that last less than 250 days.

### Part 3b

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

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

# probability that natural pregnancy last less that 260 and 280 days  
pnorm(280, mean = 268, sd = 16) - pnorm(260, mean = 268, sd = 16)

## [1] 0.4648351

## There is a 46.5% chance that a pregnancy will last 260 to 280 days.

### Part 3c

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

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

# the longest 10% of pregnancies are  
qnorm(.9, mean = 268, sd = 16)

## [1] 288.5048

## That is the longest 10% of pregnancies are 288.5048 days or greater.

## 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 -|-|-|-|-|-|-|-|-|-|-|-

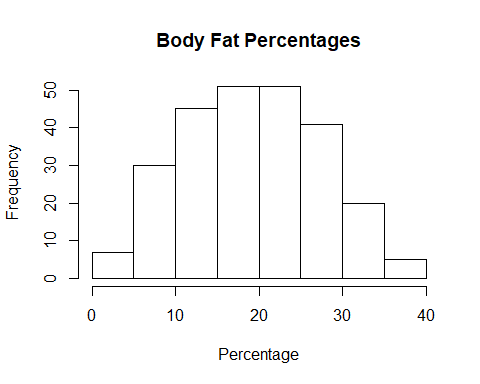
bfat <- read.csv("bodyFat.csv")

### 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$bodyFat,xlab="Percentage",main = "Body Fat Percentages")



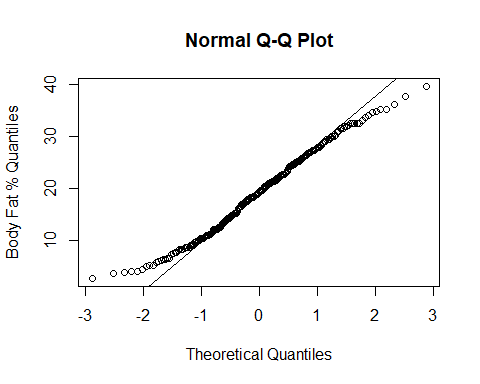
By inspection of the histogram, body fat percentage does appear to be normally distributed with mean near 20%.

### 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$bodyFat, ylab="Body Fat % Quantiles")   
 qqline(bfat$bodyFat)  
}



I think since the points curve above the line at the left and below the line at the right, the body fat percentage distribution has very short tails compared to a normal distribution. This is evidence that the body fat percentage data may not be normally distributed.