Week 1 HW Submission

*type your name here*

*enter today’s date*

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

# extract the cholesterol level of the 40 women and assign it to the variable fs  
head(HealthExam)

## Sex AgeGroup Height Weight Waist Pulse SysBP DiasBP BodyMass Cholesterol  
## 1 M 36 to 64 71.0 162.4 77.0 56 116 63 23.4 613  
## 2 M 36 to 64 72.9 174.8 88.0 56 109 74 23.8 339  
## 3 M 65+ 74.1 191.1 91.8 56 112 61 25.2 250  
## 4 M 36 to 64 64.7 119.5 75.7 56 125 65 20.7 277  
## 5 M 36 to 64 72.1 189.1 95.0 56 120 68 26.3 649  
## 6 M 65+ 69.7 214.5 107.5 56 125 69 32.1 288  
## Leg Region  
## 1 43.1 West  
## 2 44.2 South  
## 3 48.4 Midwest  
## 4 42.6 West  
## 5 44.9 Northeast  
## 6 42.8 Midwest

fs <- subset(HealthExam$Cholesterol, HealthExam$Sex=='F')  
mean(fs)

## [1] 240.875

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

# Summarize dataset and apply sd to dataset to get standard deviation  
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

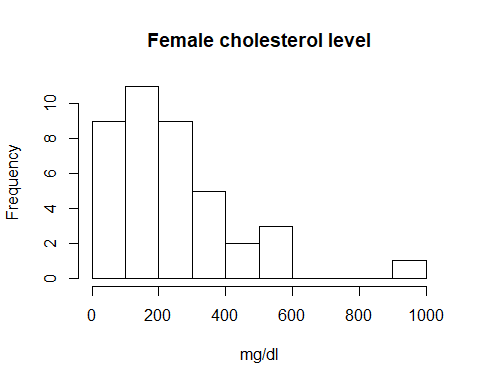
Based on the Summary Stats the data does appear to be Right-Skewed.

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



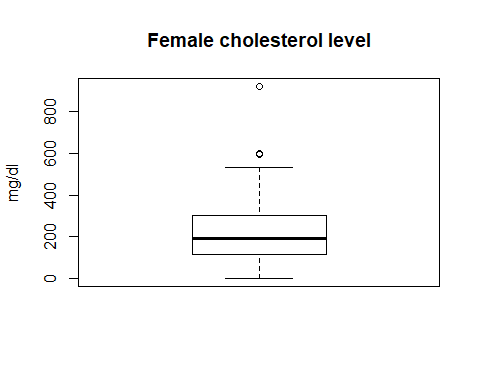
No it does not appear to be normally distributed. It appears to be left-skewed

### Part 1d

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

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

# oxplot for 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  
# IQR(fs) gives the interquartile range  
out\_cut <- IQR(fs)\*1.5  
out\_cut

## [1] 274.125

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

# Z score calculation  
z <- (max(fs) - mean(fs))/sd(fs)  
z

## [1] 3.651556

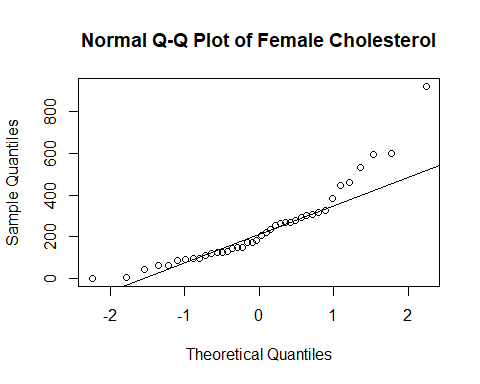
The maximum femaale cholesterol level does seem unusually large as it is more than 3 standard deviations from the mean

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

# normal probability plot for fs  
#plot(fs)  
  
qqnorm(fs,main = "Normal Q-Q Plot of Female Cholesterol",  
 xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")  
qqline(fs)



The sample does not seem to come from a normal distribution as there are measures above the line at the beginning and at the end as well. The values do not appear to be normally distributed based on the plot

## 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  
testScores <- read.csv("testScores.csv")  
head(testScores)

## score instructor  
## 1 66 A  
## 2 70 A  
## 3 75 A  
## 4 54 A  
## 5 66 A  
## 6 71 A

# summarize the scores for each instructor  
summary(testScores$score[testScores$instructor=="A"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 52.00 61.00 65.50 65.73 69.75 79.00

summary(testScores$score[testScores$instructor=="B"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 54.00 70.25 74.00 72.63 76.75 85.00

summary(testScores$score[testScores$instructor=="C"])

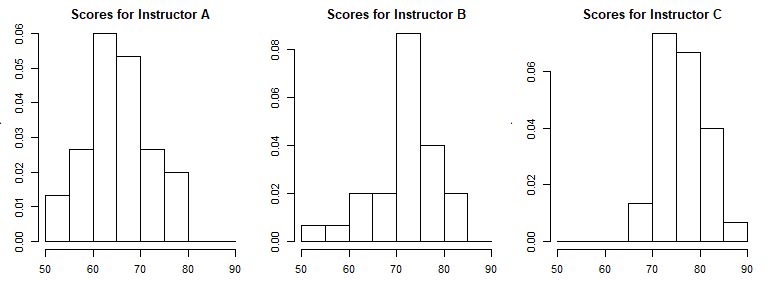
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 69.00 72.25 76.00 76.53 79.00 87.00

# a more advanced version of this summary by instructor would be to use tapply() to loop over the factors.  
# uncomment the next line to try it  
tapply(testScores$score, testScores$instructor, summary)

## $A  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 52.00 61.00 65.50 65.73 69.75 79.00   
##   
## $B  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 54.00 70.25 74.00 72.63 76.75 85.00   
##   
## $C  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 69.00 72.25 76.00 76.53 79.00 87.00

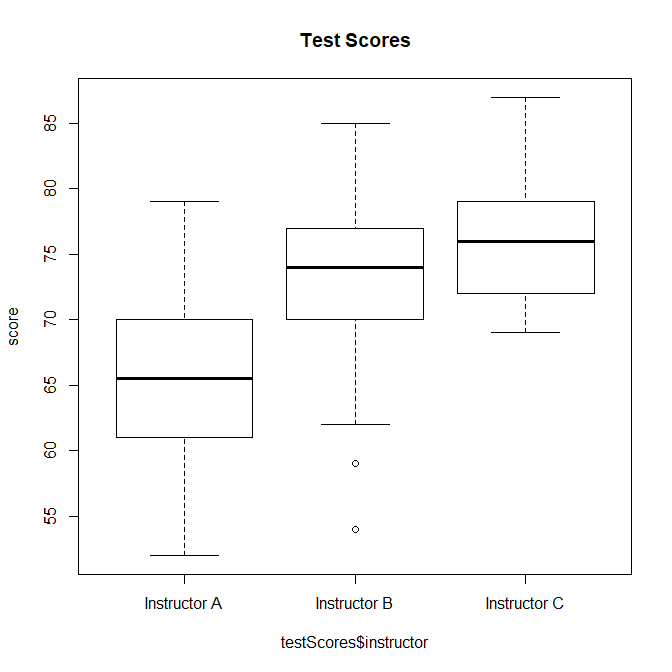
# looks like the test scores across all three instructors range from 52 to 87, to make it easier to compare the distributions lets use the same bins for all three histograms  
bins <- seq(50,90,by=5)

# 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(testScores$score[testScores$instructor=="A"],main="Scores for Instructor A",breaks = bins,xlab="score", prob=TRUE)  
hist(testScores$score[testScores$instructor=="B"],main="Scores for Instructor B",breaks = bins,xlab="score", prob=TRUE,ylab="")  
hist(testScores$score[testScores$instructor=="C"],main="Scores for Instructor C",breaks = bins,xlab="score", prob=TRUE)



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

# make side-by-side boxplots to make comparisons easier  
boxplot(testScores$score~testScores$instructor,names=c("Instructor A","Instructor B","Instructor C"),ylab="score",main="Test Scores")



### PART 2 A

# 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 scores for each instructor  
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

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

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

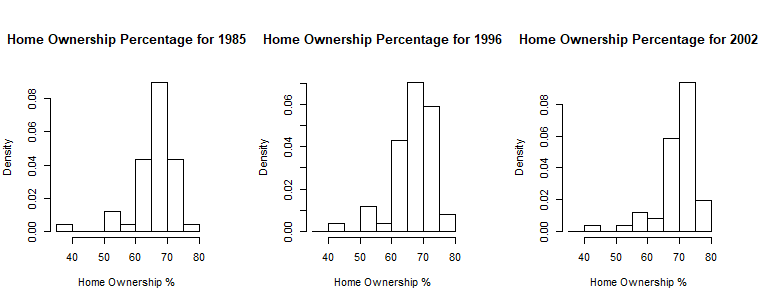
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 44.10 67.25 70.20 69.45 73.50 77.30

# a more advanced version of this summary by instructor would be to use tapply() to loop over the factors.  
# uncomment the next line to try it  
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

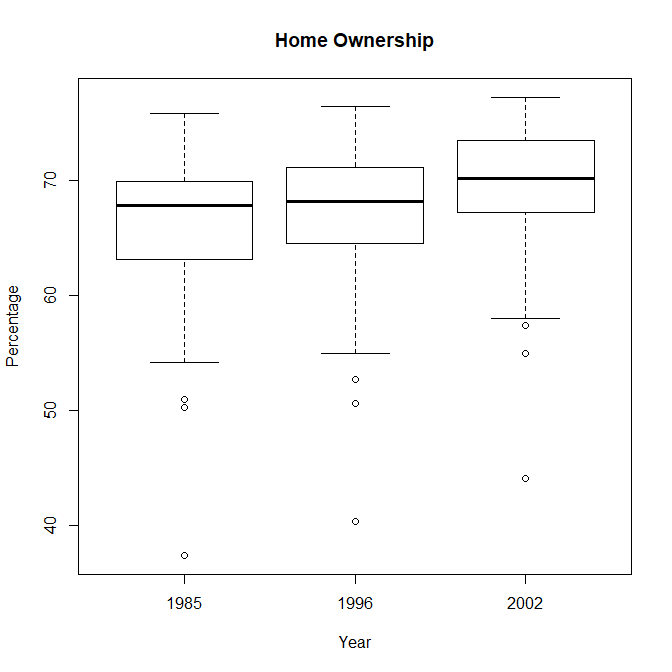
# looks like the home ownership percentage across all threeyears range from 37 to 77, to make it easier to compare the distributions lets use the same bins for all three histograms  
bins <- seq(35,80,by=5)

# 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(4,4,6,4))  
hist(homes$PercentOwned[homes$Year=="year1985"],main="Home Ownership Percentage for 1985",breaks = bins,xlab="Home Ownership %", prob=TRUE)  
hist(homes$PercentOwned[homes$Year=="year1996"],main="Home Ownership Percentage for 1996",breaks = bins,xlab="Home Ownership %", prob=TRUE)  
hist(homes$PercentOwned[homes$Year=="year2002"],main="Home Ownership Percentage for 2002",breaks = bins,xlab="Home Ownership %", 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("1985","1996","2002"),ylab="Percentage",xlab="Year",main="Home Ownership")



#End PArt 2A

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

Yes there does appear to be a general trend in home ownership between the 3 years.It seems to be increasing For similarities, each year has 3 outliers below the mean. The means of the 3 years are fairly close together. Same with the top rates as well. For differences, the outliers for 1996 and 2002 seem to be spaced similar distacne apart but for 1985 they are much closer together.

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

z <- (250-268)/16  
  
pnorm(z)

## [1] 0.1302945

### Part 3b

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

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

p260 = 260  
p280 = 280  
  
p1 <- (p260 -268)/16  
pn1 <- pnorm(p1)  
   
  
  
p2 <- (p280-268)/16  
pn2 <- pnorm(p2)  
   
  
  
panswr <- pn2 - pn1  
  
panswr

## [1] 0.4648351

### Part 3c

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

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

# insert your code here.  
z10 <- pnorm(.1)  
  
qnorm(.9,mean=268,sd =16,lower.tail=TRUE)

## [1] 288.5048

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

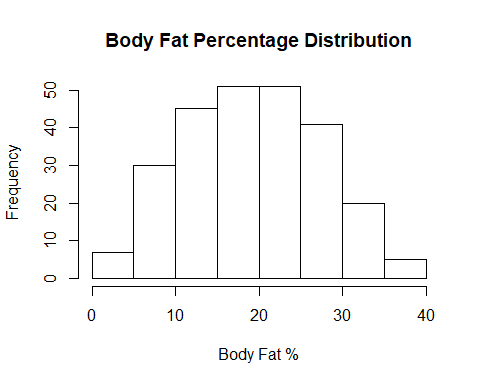
# insert your code here.  
bfat\_df <- read.csv("bodyFat.csv")  
bfat <- bfat\_df[["bodyFat"]]

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

# Make a histogram of the bodyfat percentages  
hist(bfat,main="Body Fat Percentage Distribution",xlab="Body Fat %")



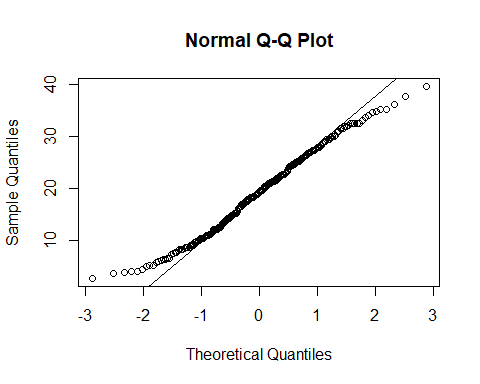
IT does appear that the data does come from a normally distributed variable. The values are higher in the middle and lower on both ends. IT you put a dsitribution curve over the top of this it would be bell shaped which indicates normali

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

# normal probability plot with a fitted line by using qqnorm() and qqline()  
qqnorm(bfat)  
qqline(bfat)



This tells me that the bodyfat data is not normally distributed. IT does not follow the line. At the beginning of the line there is data above the line and the data points start to the left of there the line begins. At the top of the line, there a points that are below the line and do not follow the line.