Week 1 HW Submission

Spencer Swartz

January 23, 2017

## 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.

## 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 variable contains the 40 women in the HealthExam dataset.  
fs <- HealthExam[which(HealthExam$Sex=='F'),]  
fs <- fs$Cholesterol  
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 -|-|-|-|-|-|-|-|-|-|-|-

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

Replace text with your written answer.

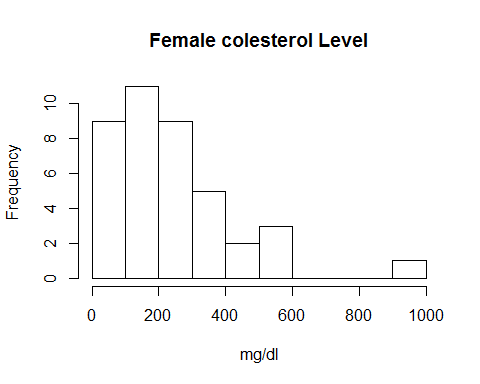
There may be some skew towards the left.

### 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, main = 'Female colesterol Level', xlab = 'mg/dl')



Replace text with your written answer.

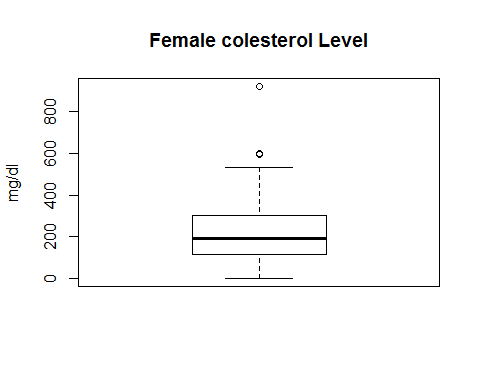
The cholesterol levels do not seem to be normal distributed since there is a skew and a hard stop at 0.

### 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, main='Female colesterol Level', ylab='mg/dl')



### 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) gives the third quartile  
# IQR(fs) gives the interquartile range  
(IQR(fs)\*1.5)+quantile(fs,.75)

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

# insert your code here.  
(max(fs)-mean(fs))/sd(fs)

## [1] 3.651556

Replace this text with your written answer.

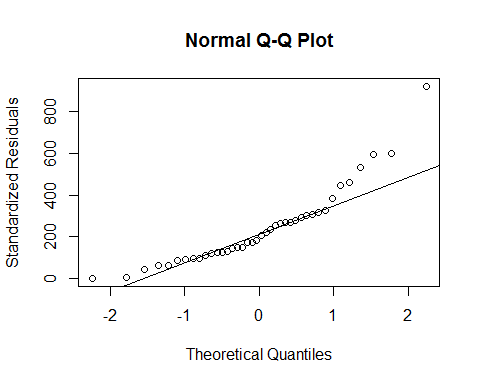
since the Z-score is greater than three we can see that the max data point is an outlier as it is more than 3 standard deviations from the mean.

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

qqnorm(fs, ylab = 'Standardized Residuals')  
qqline(fs)



It looks like it may be Normally distributed other than the end of the lines where the outliers seem to be effecting it a little bit.

## 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 (re-scale 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 -|-|-|-|-|-|-|-|-|-|-|-

# 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)

# load the data and look at it  
homes <- read.csv("homes.csv")  
head(homes)

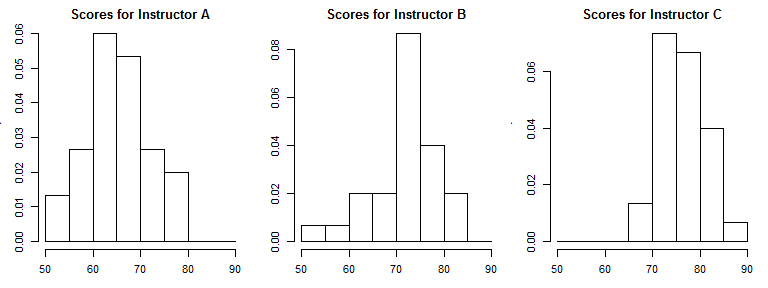
## PercentOwned Year  
## 1 70.4 year1985  
## 2 61.2 year1985  
## 3 64.7 year1985  
## 4 66.6 year1985  
## 5 54.2 year1985  
## 6 63.6 year1985

# summarize the scores for each year  
#summary(homes$PercentOwned[homes$Year=="year1985"])  
#summary(homes$PercentOwned[homes$Year=="year1996"])  
#summary(homes$PercentOwned[homes$Year=="year2002"])  
  
# a more advanced version of this summary by year 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

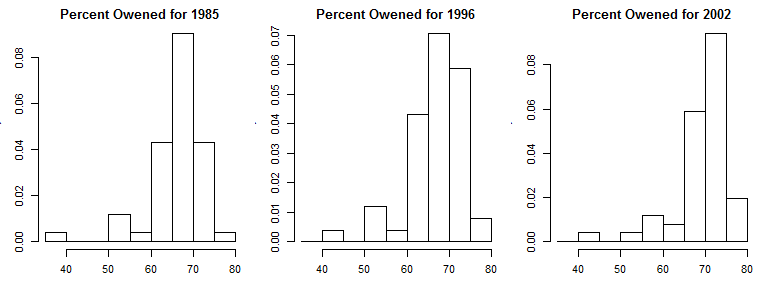
# looks like the test scores across all three year range from 37.4 to 77.3, to make it easier to compare the distributions lets use the same bins for all three histograms  
homebins <- seq(35,80,by=5)

# 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(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)



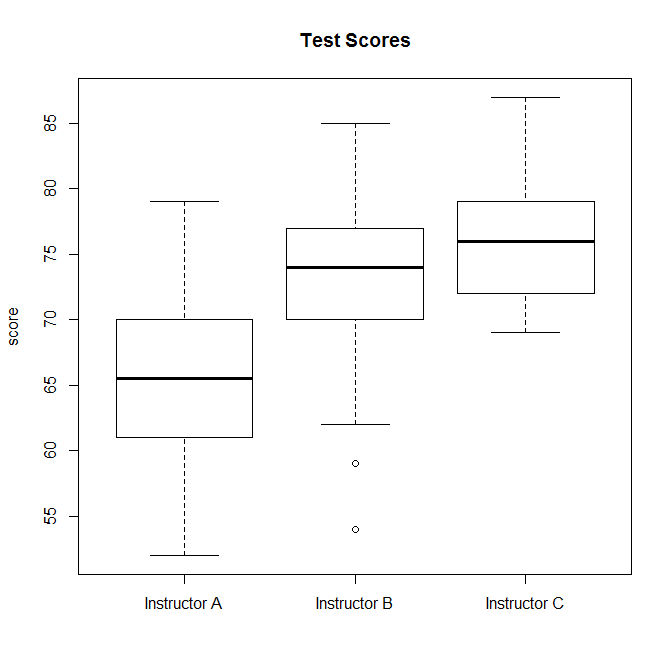
# reset to one plot  
par(mfrow=c(1,1))

# 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 Owened for 1985",breaks = homebins,xlab="%owned", prob=TRUE)  
hist(homes$PercentOwned[homes$Year=="year1996"],main="Percent Owened for 1996",breaks = homebins,xlab="%owned", prob=TRUE)  
hist(homes$PercentOwned[homes$Year=="year2002"],main="Percent Owened for 2002",breaks = homebins,xlab="%owned", prob=TRUE)

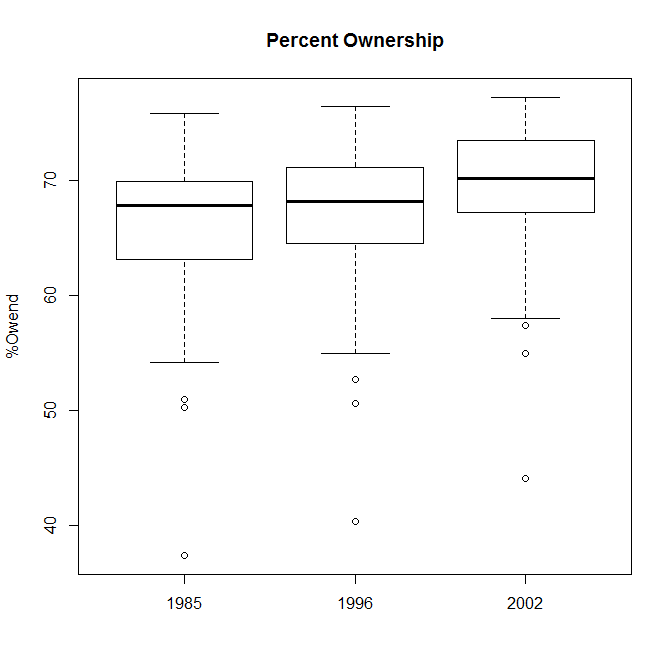


# reset to one plot  
par(mfrow=c(1,1))

# 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")



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



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

In General it seems that home ownership is increasing as time goes on. the biggest difference was between 1996 and 2002 as the median ownership percent jumped 2 percentage points.

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

# insert your code here.  
1-pnorm(250, mean = 268, sd = 16, lower.tail = FALSE)

## [1] 0.1302945

### Part 3b

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

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

# insert your code here.  
pnorm(280, mean = 268, sd = 16, lower.tail = TRUE)-pnorm(260, mean = 268, sd = 16, lower.tail = TRUE)

## [1] 0.4648351

### Part 3c

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

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

# insert your code here.  
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 -|-|-|-|-|-|-|-|-|-|-|-

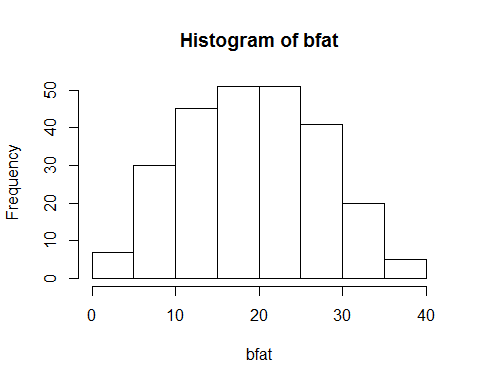
# insert your code here.  
bodyFat <- read.csv("bodyFat.csv")  
bfat <- bodyFat$bodyFat

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

# insert your code here.  
hist(bfat)



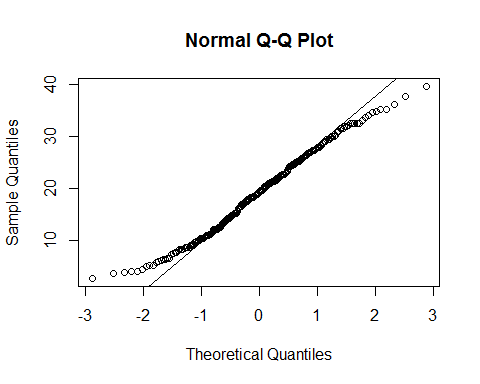
Yes this looks to be a normal distribution and it would be easy to fit a normal curve to it.

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

# insert your code here.  
qqnorm(bfat)  
qqline(bfat)



It looks like we may have fat tails to the distribution. This means that the values do not taper of as quickly as a typical normal distribution.