Week 1 HW Submission

Josh Jarvey

5/21/2020

## 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.

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 -|-|-|-|-|-|-|-|-|-|-|-

#return the row of the observations that are Female from the HealthExam dataset - stored into fs.  
fs = HealthExam[HealthExam$Sex=="F",]  
  
 #calculate the mean cholesterol of the new (female) dataset.   
mean(fs$Cholesterol)

## [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 the cholesterol variable for the female dataset  
summary(fs$Cholesterol)

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

#calculate the standard deviation for the cholesterol variable  
sd(fs$Cholesterol)

## [1] 185.9824

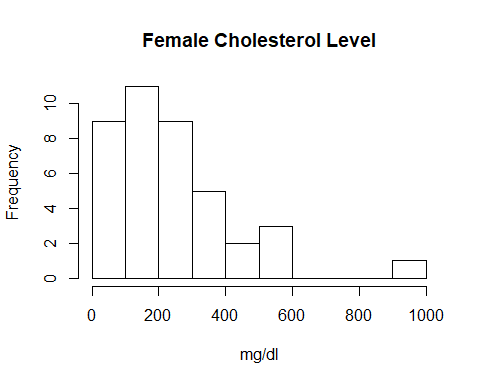
Yes the data appears to be skewed to the right, as the mean is greater than the median.

### 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 1c -|-|-|-|-|-|-|-|-|-|-|-

#create a histogram of female cholesterol levels  
hist(fs$Cholesterol, xlab = "mg/dl", main = "Female Cholesterol Level")



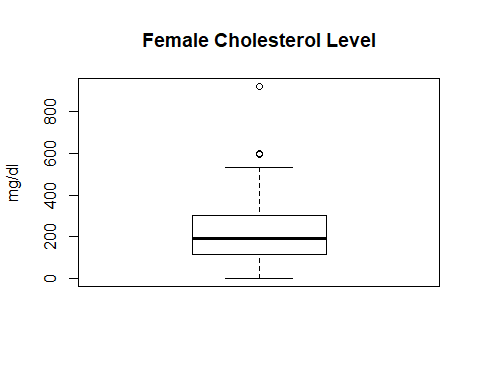
The shape of the histogram further indicates my suspicion that the data is right-skewed. As for normality, the data does not appear to be bell-shaped. Furthermore, there will be no observations below 0, since an individual cannot have a negative cholesterol level. Could this be a Poisson distribution?

### Part 1d

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

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

#create a boxplot for female cholesterol levels  
boxplot(fs$Cholesterol, 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  
quantile(fs$Cholesterol,0.75) + 1.5\*IQR(fs$Cholesterol)

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

#z-score = (x-mean)/sd  
(max(fs$Cholesterol) - mean(fs$Cholesterol)) / sd(fs$Cholesterol)

## [1] 3.651556

#calculate the probability of a z-score of this magnitude from a normal distribution (using either tail).  
 #Multiply the probability by the 40 observations, and review the expected number of observations with this z-score.  
prob=2\*pnorm(-3.65)  
prob

## [1] 0.0002622403

prob\*40

## [1] 0.01048961

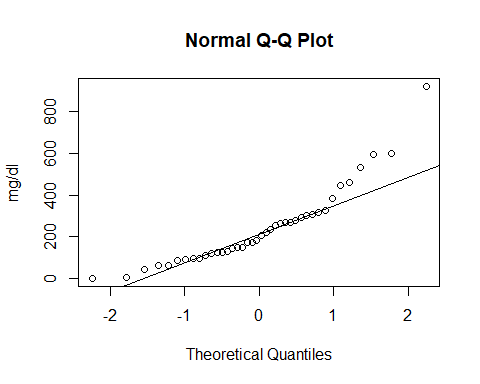
The 920 mg/dl cholesterol observation does appear unusually large, given that it is ~3.65 standard deviations away from the mean. If this were a normal distribution, a z-score this large in probability terms would be ~0.00026. So in 40 observations from a normal distribution we should expect 40\*0.00026 = ~0.01 observations that have this z-score. To have one is very rare.

### Part 1g

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

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

#Create a QQ Plot using the cholesterol data.  
qqnorm(fs$Cholesterol, ylab="mg/dl")  
qqline(fs$Cholesterol)



The data does not appear to come from a normal distribution because the tails on the QQ plot tend to shift upward away from the theoritical normal distribution (the line). The lower portion of the tail indicates skew to the right (since its above the line), and the top tail indicates a longer than normal tail in the distribution (since its above the line) - further indicating a right-skew dataset that does not come from the normal distribution.

## 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.

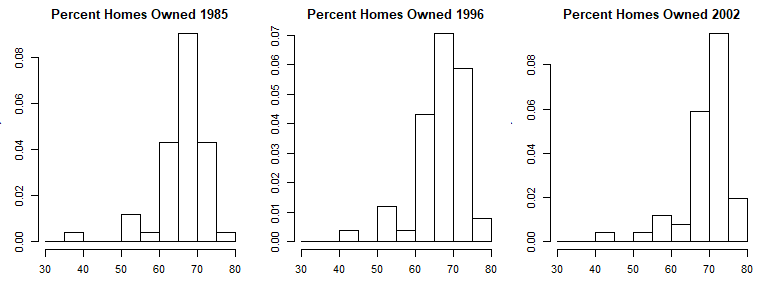
### -|-|-|-|-|-|-|-|-|-|-|- Answer 2a -|-|-|-|-|-|-|-|-|-|-|-

homeownership <- read.csv("homes.csv")  
  
tapply(homeownership$PercentOwned, homeownership$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

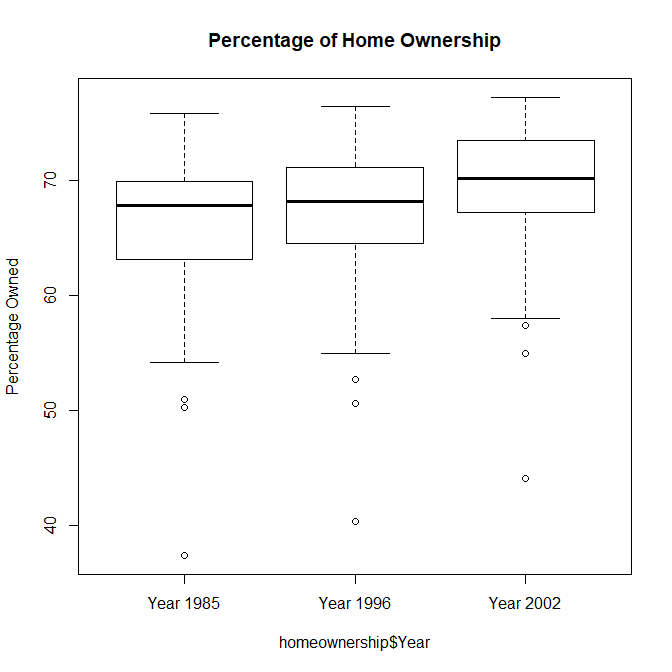
bins = seq(30,80,by=5)

# 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(homeownership$PercentOwned[homeownership$Year=="year1985"],main="Percent Homes Owned 1985",breaks = bins,xlab="Percent", prob=TRUE)  
hist(homeownership$PercentOwned[homeownership$Year=="year1996"],main="Percent Homes Owned 1996",breaks = bins,xlab="Percent", prob=TRUE,ylab="")  
hist(homeownership$PercentOwned[homeownership$Year=="year2002"],main="Percent Homes Owned 2002",breaks = bins,xlab="Percent", prob=TRUE)



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

# make side-by-side boxplots to make comparisons easier  
boxplot(homeownership$PercentOwned~homeownership$Year,names=c("Year 1985","Year 1996","Year 2002"),ylab="Percentage Owned",main="Percentage of Home 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 -|-|-|-|-|-|-|-|-|-|-|-

The home ownership amongst all groups (year) are left-skewed, with multiple outliers in that direction. Additionally, there does appear to be a slight positive linear increase to ownership rate, indicating that as the years go by more individuals are owning their home (vs. renting).

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

#finding the probability of 250 days, 268 mean, 16 std.  
pnorm(250,268,16)

## [1] 0.1302945

### Part 3b

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

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

#first find the probability of 280 days, then subtract out the probability of 260 days. This remaining "area" is our between prob.   
pnorm(280,250,16) - pnorm(260,250,16)

## [1] 0.2355892

### Part 3c

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

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

#using qnorm to find the 90th quantile, since this is our cutoff to answer the question. The longest 10% are at least 270.5 days or greater.   
qnorm(.9,250,16)

## [1] 270.5048

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

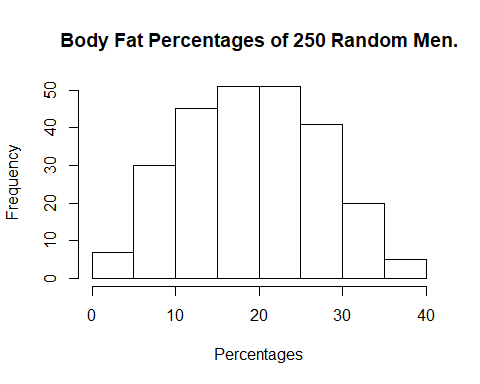
#read in the body fat data.  
bf = read.csv("bodyfat.csv")  
 #extract the data into a vector.  
bfat = bf$bodyFat

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

#create a histogram of the body fat data.  
hist(bfat, main = "Body Fat Percentages of 250 Random Men.", xlab = "Percentages")



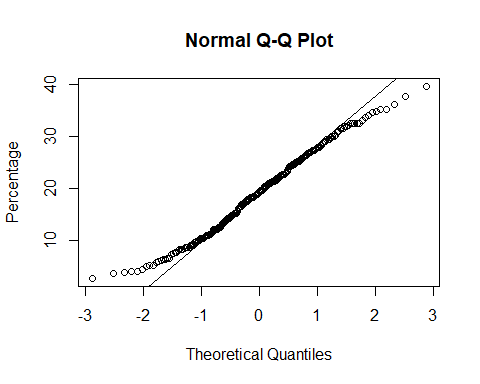
Yes, the data does appear to come from a normal distribution. It displays a pretty strong “bell-shaped” curve, and there is no significant skew in either direction.

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

#create a QQ plot of the Body Fat data.  
qqnorm(bfat, ylab = "Percentage")  
qqline(bfat)



The bottom half of the graph shows that the observed data points are greater than what would be expected, and the top half of the graph shows the observed data points are less than the expected theoritical distribution. This indicates underdispersion in the sample data set, as there doesnt appear to be enough variability vs. what would be expected from a normal distribution. (Thanks Dr. Baumann for helping me further interpret this plot!)