

9 Additional Graphs Homework

Amin Baabol

07/27/2020

General Instructions

There are six exercises below. You are required to provide five solutions, with the same options for choosing languages as with the last exercise. You can provide solutions in two languages for one exercise only (for example, Ex. 1,2,3,5 in R and Ex. 1 in SAS is acceptable, Ex. 1,2,3 in SAS and Ex. 1,2 in R is not)

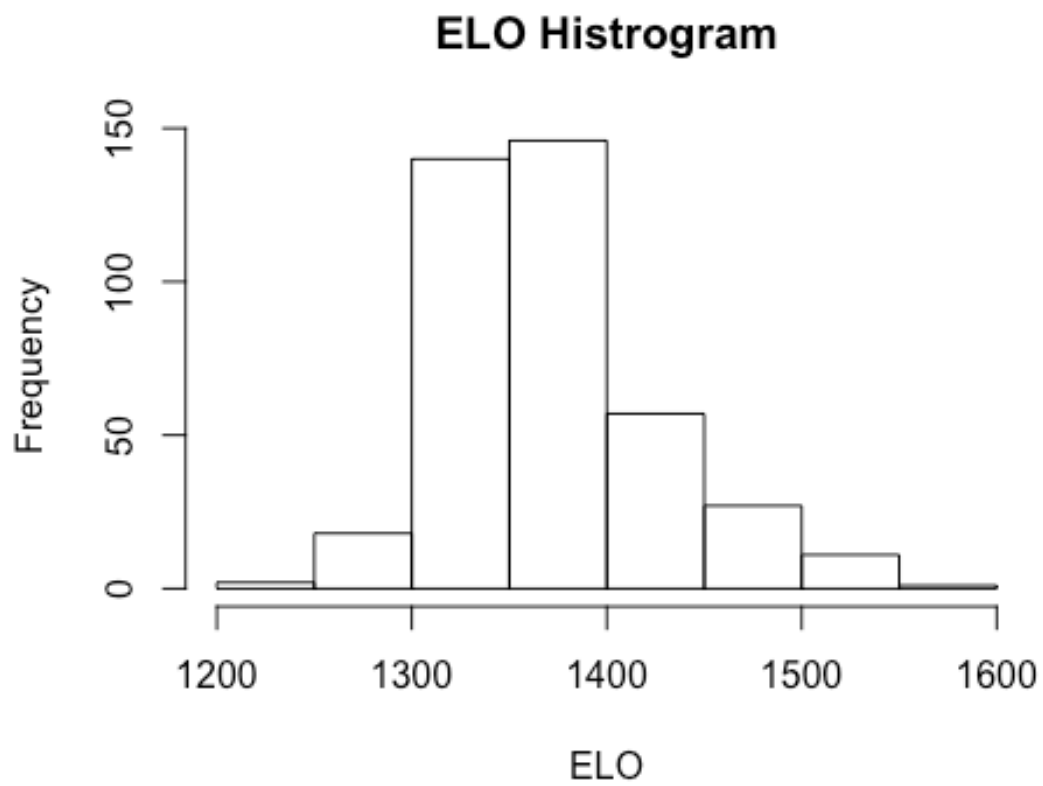
For this exercise, you may use whatever graphics library you desire.

Exercise 1.

Load the `ncaa2018.csv` data set and create histograms, QQ-norm and box-whisker plots for ELO. Add a title to each plot, identifying the data.

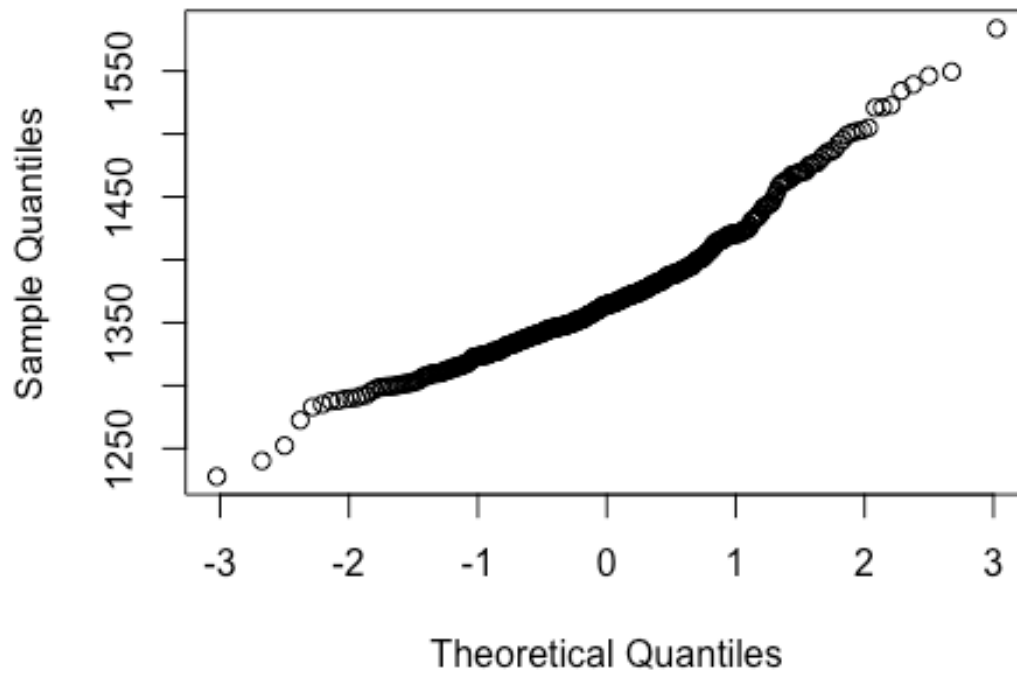
```
#Loading ncaa2018.csv data set
ncaa2018csv =
read.table("~/Desktop/work/GradSchool/Summer2020/STATS600/week7/ncaa2018.csv"
,header = T,sep = ",")

#Histogram
hist(ncaa2018csv$ELO, xlab = "ELO", main = "ELO Histogram")
```

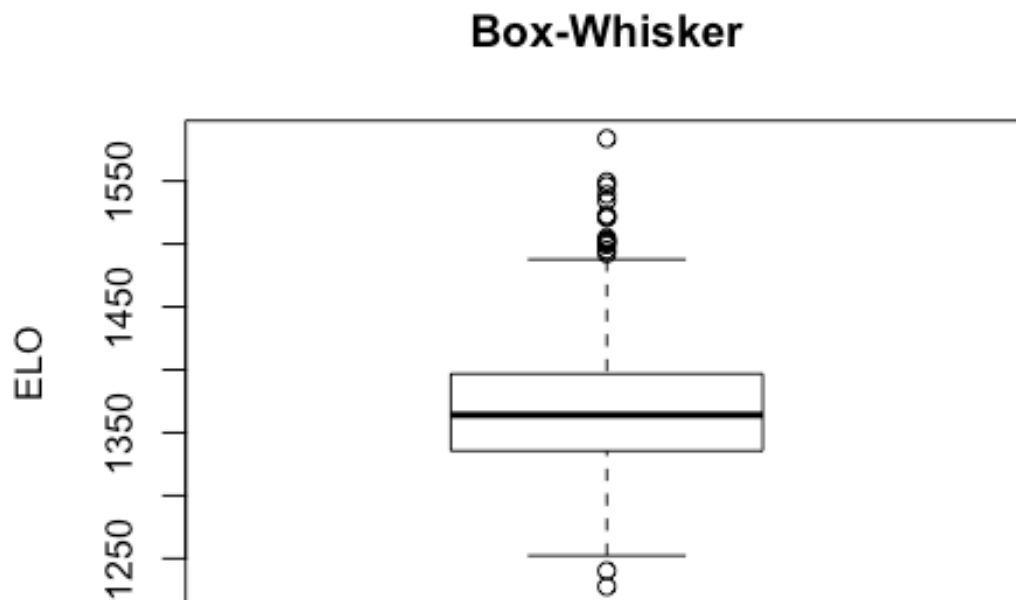


```
# QQ-Norm  
qqnorm(ncaa2018csv$ELO, main = "Normal QQ plot")
```

Normal QQ plot



```
# Box-Whisker  
boxplot(ncaa2018csv$ELO, main = "Box-Whisker", ylab = "ELO")
```



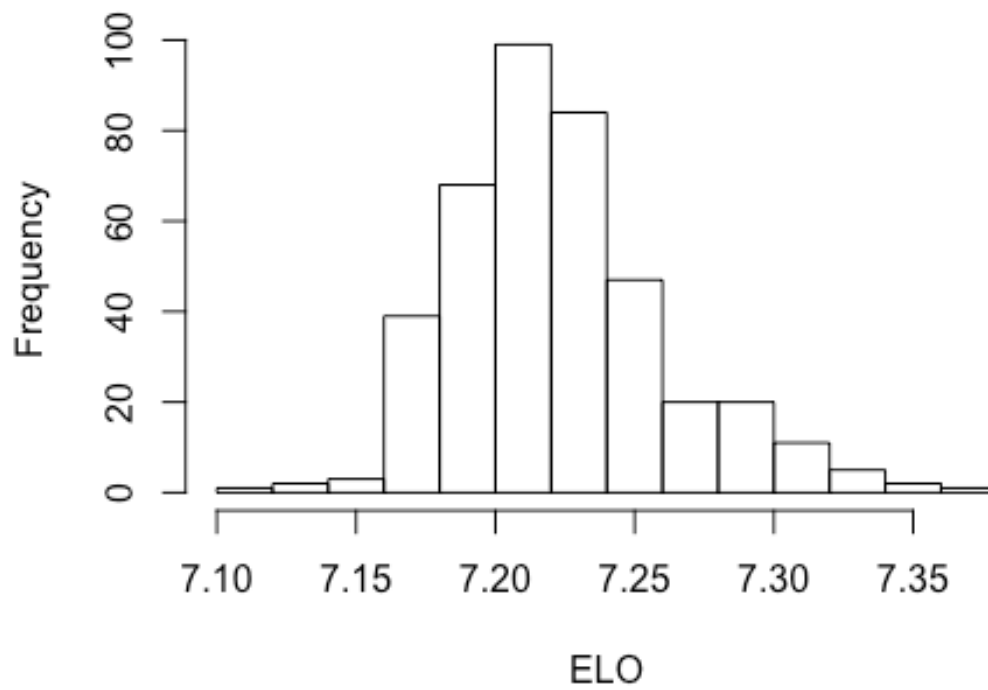
Part b

A common recommendation to address issues of non-normality is to transform data to correct for skewness. One common transformation is the log transform.

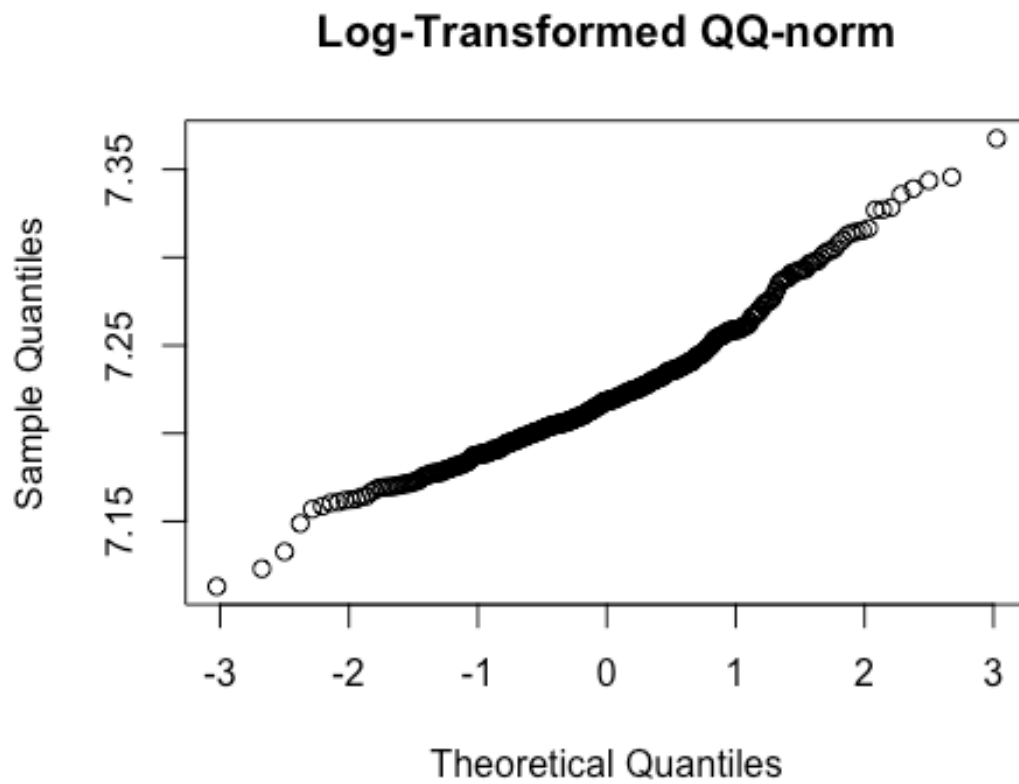
Transform ELO to $\log(\text{ELO})$ and produce histograms, box-whisker and qqnorm plots of the transformed values. Are the transformed values more or less skewed than the original? You might calculate skewness and kurtosis values as in Homework 6, Exercise 2.

```
#Histogram  
hist(log(ncaa2018csv$ELO), xlab = "ELO", main = "Log-Transformed Histogram")
```

Log-Transformed Histogram

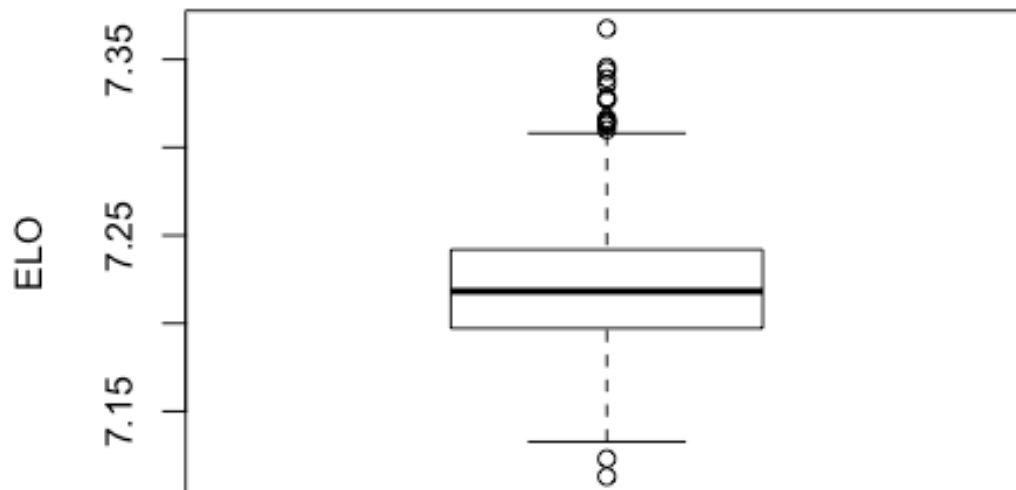


```
# QQ plot  
qqnorm(log(ncaa2018csv$ELO), main = "Log-Transformed QQ-norm")
```



```
# box-whisker plot  
boxplot(log(ncaa2018csv$ELO), main = "Log-Transformed Box-whisker", ylab =  
"ELO")
```

Log-Transformed Box-whisker



Exercise 2.

Review Exercise 2, Homework 6, where you calculated skewness and kurtosis. The reference for this exercise,

<https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm>,

The following example shows histograms for 10,000 random numbers generated from a normal, a double exponential, a Cauchy, and a Weibull distribution.

We will reproduce the histograms for these samples, and add qqnorm and box-whisker plots.

Part a

Use the code below from lecture to draw 10000 samples from the normal distribution.

```
norm.sample <- rnorm(10000, mean=0, sd=1)
```

Look up the corresponding `r*` functions in R for the Cauchy distribution (use `location=0`, `scale=1`), and the Weibull distribution (use `shape = 1.5`). For the double exponential, use you can use the `*laplace` functions from the `rmutil` library, or you can use `rexp(10000) - rexp(10000)`

Draw 10000 samples from each of these distributions. Calculate skewness and kurtosis for each sample. You may use your own function, or use the moments library.

```
cauchy.distro<- rcauchy(1:1000, location = 0, scale = 1)
weibull.distro <- rweibull(1:1000, shape = 1.5)
rexp10k <- rexp(1000) - rexp(1000)

#install.packages(moments)
library(moments)

#Normal Distribution Skewness and Kurtosis
skewness1 <- round(skewness(norm.sample, na.rm = TRUE),2)
kurtosis1 <- round(kurtosis(norm.sample, na.rm = TRUE),2)
skewness1

## [1] 0.01

kurtosis1

## [1] 2.95

#Cauchy Distribution Skewness and Kurtosis
skewness2 <- round(skewness(cauchy.distro, na.rm = TRUE),2)
kurtosis2 <- round(kurtosis(cauchy.distro, na.rm = TRUE),2)
skewness2

## [1] -19.34

kurtosis2

## [1] 444

#Weibull Distribution Skewness and Kurtosis

skewness3 <- round(skewness(weibull.distro, na.rm = TRUE),2)
kurtosis3 <- round(kurtosis(weibull.distro, na.rm = TRUE),2)
skewness3

## [1] 0.92

kurtosis3

## [1] 3.92

#Exponential Distribution Skewness and Kurtosis
skewness4 <- round(skewness(rexp10k, na.rm = TRUE),2)
kurtosis4 <- round(kurtosis(rexp10k, na.rm = TRUE),2)
skewness4

## [1] -0.25

kurtosis4
```

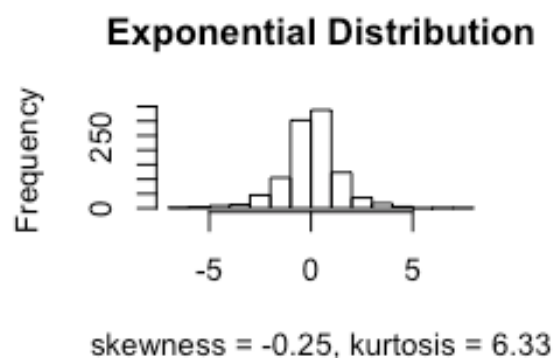
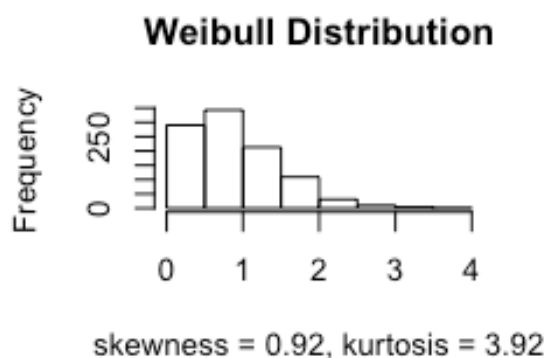
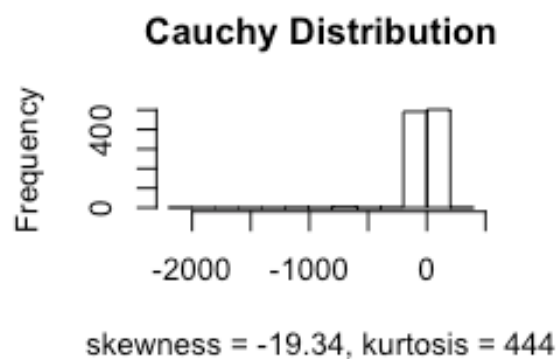
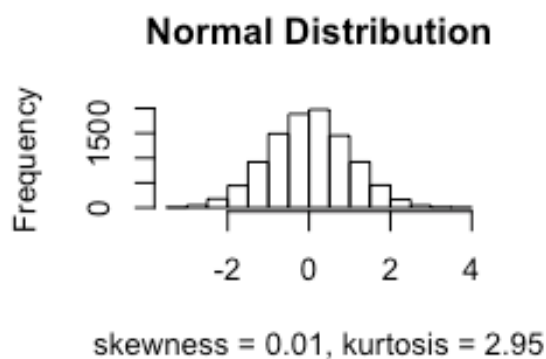


```
## [1] 6.33
```

Part b

Plot the histograms for each distribution. Use `par(mfrow=c(2,2))` in your code chunk to combine the four histograms in a single plot. Add titles to the histograms indicating the distribution. Set the x-axis label to show the calculated skewness and kurtosis, i.e. `skewness = ####, kurtosis = ####`

```
par(mfrow=c(2,2))
hist(norm.sample, xlab = paste0("skewness = ", skewness1 , ", kurtosis = ",
kurtosis1), main = "Normal Distribution")
hist(cauchy.distro, xlab = paste0("skewness = ", skewness2 , ", kurtosis = ",
kurtosis2), main = "Cauchy Distribution")
hist(weibull.distro, xlab = paste0("skewness = ", skewness3 , ", kurtosis = ",
", kurtosis3), main = "Weibull Distribution")
hist(rexp10k, xlab = paste0("skewness = ", skewness4 , ", kurtosis = ",
kurtosis4), main = "Exponential Distribution")
```



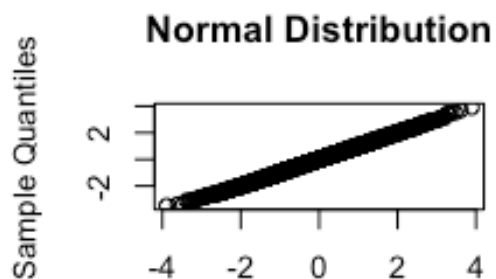
Part c

Repeat Part b, but with QQ-norm plots.

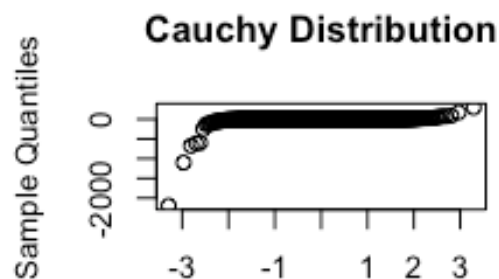
```

par(mfrow=c(2,2))
qqnorm(norm.sample, xlab = paste0("skewness = ", skewness1 , ", kurtosis = ",
kurtosis1), main = "Normal Distribution")
qqnorm(cauchy.distro, xlab = paste0("skewness = ", skewness2 , ", kurtosis = ",
", kurtosis2), main = "Cauchy Distribution")
qqnorm(weibull.distro, xlab = paste0("skewness = ", skewness3 , ", kurtosis = ",
", kurtosis3), main = "Weibull Distribution")
qqnorm(rexp10k, xlab = paste0("skewness = ", skewness4 , ", kurtosis = ",
kurtosis4), main = "Exponential Distribution")

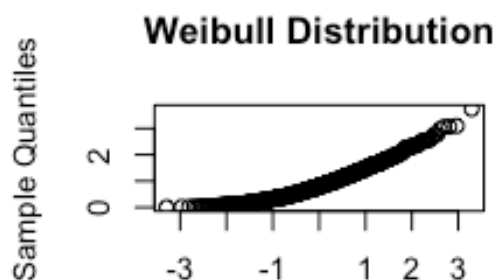
```



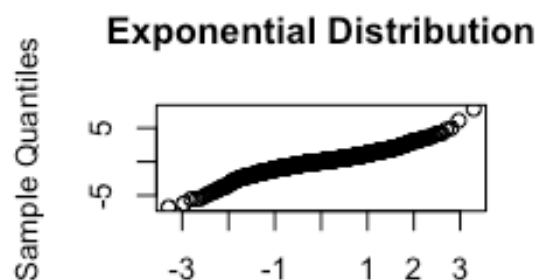
skewness = 0.01, kurtosis = 2.95



skewness = -19.34, kurtosis = 444



skewness = 0.92, kurtosis = 3.92



skewness = -0.25, kurtosis = 6.33

Part d

Repeat Part b, but with box-whisker plots.

```

par(mfrow=c(2,2))
boxplot(norm.sample, xlab = paste0("skewness = ", skewness1 , ", kurtosis = ",
", kurtosis1), main = "Normal Distribution")
boxplot(cauchy.distro, xlab = paste0("skewness = ", skewness2 , ", kurtosis = ",
", kurtosis2), main = "Cauchy Distribution")
boxplot(weibull.distro, xlab = paste0("skewness = ", skewness3 , ", kurtosis = ",
", kurtosis3), main = "Weibull Distribution")
boxplot(rexp10k, xlab = paste0("skewness = ", skewness4 , ", kurtosis = ",
kurtosis4), main = "Exponential Distribution")

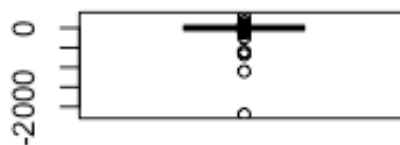
```

Normal Distribution



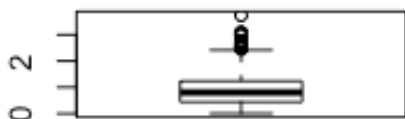
skewness = 0.01, kurtosis = 2.95

Cauchy Distribution



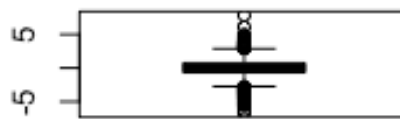
skewness = -19.34, kurtosis = 444

Weibull Distribution



skewness = 0.92, kurtosis = 3.92

Exponential Distribution



skewness = -0.25, kurtosis = 6.33

Hints for SAS. If you create the samples in IML, use

```
Normal = j(1, 10000, .);  
call randgen(Normal, "NORMAL", 0, `);
```

You can generate samples in the data step using

```
do i = 1 to 10000;  
  Normal = rand('NORMAL',0,1);  
  output;  
end;
```

RAND doesn't provide a Laplace option, but you can create samples from this distribution by

```
rand('EXPONENTIAL')-rand('EXPONENTIAL');
```

To group multiple plots, use

```
ods graphics / width=8cm height=8cm;  
ods layout gridded columns=2;  
ods region;  
... first plot
```

```
ods region;  
... second plot
```

```
ods layout end;
```

You might need to include

```
ods graphics off;  
  
ods graphics on;  
ODS GRAPHICS / reset=All;
```

to return the SAS graphics output to normal.

Exercise 3.

We will create a series of graphs illustrating how the Poisson distribution approaches the normal distribution with large λ . We will iterate over a sequence of λ , from 2 to 64, doubling λ each time. For each ' λ ' draw 1000 samples from the Poisson distribution.

Calculate the skewness of each set of samples, and produce histograms, QQ-norm and box-whisker plots. You can use `par(mfrow=c(1,3))` to display all three for one λ in one line. Add `lambda=##` to the title of the histogram, and `skewness=##` to the title of the box-whisker plot.

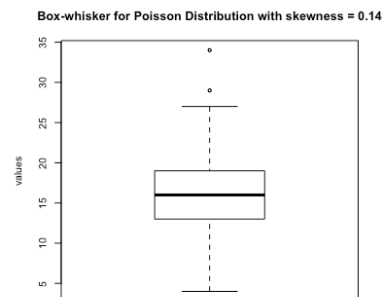
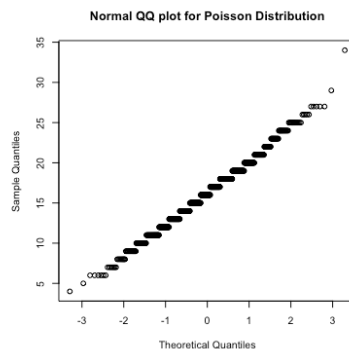
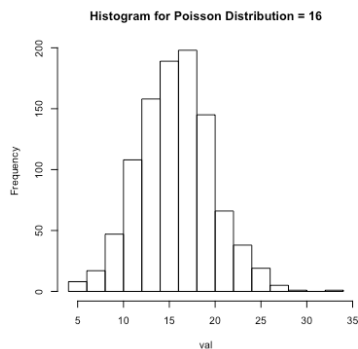
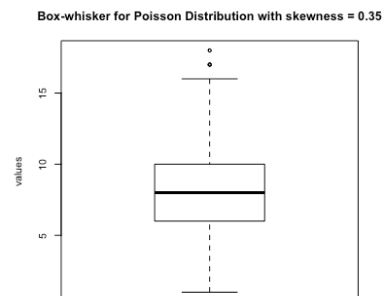
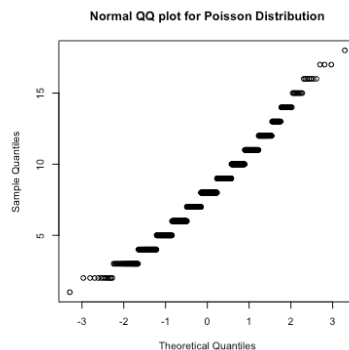
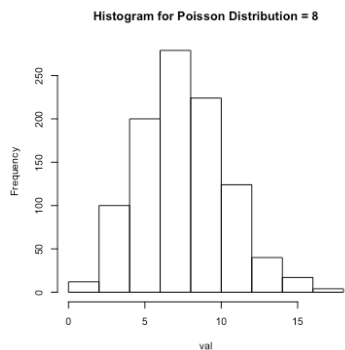
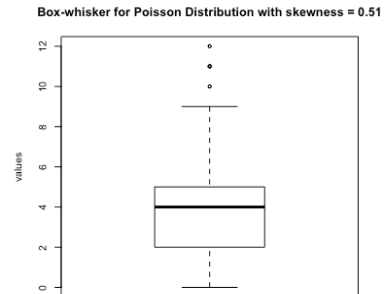
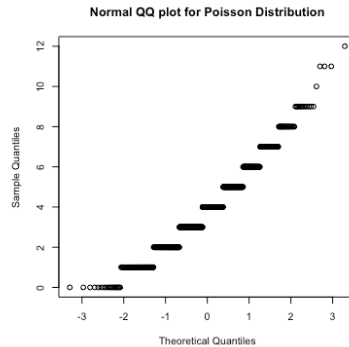
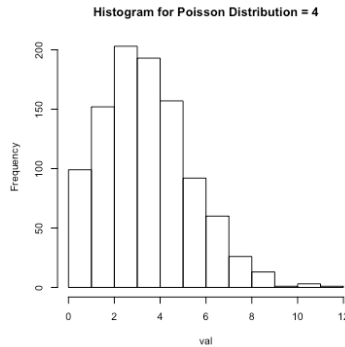
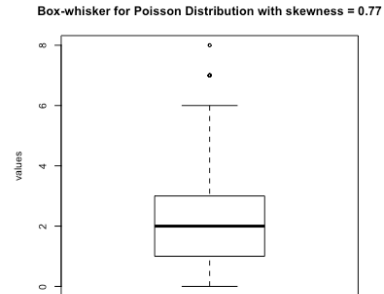
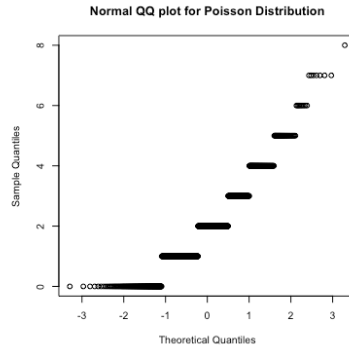
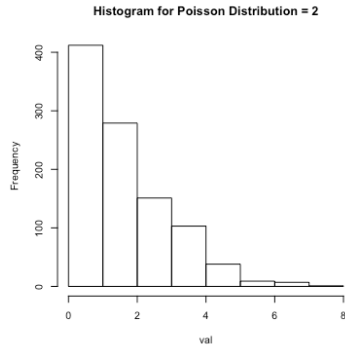
Part b.

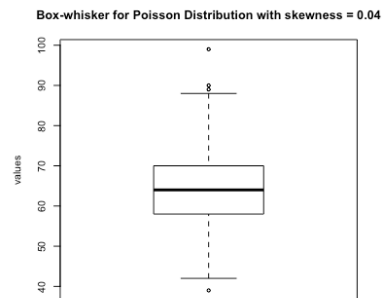
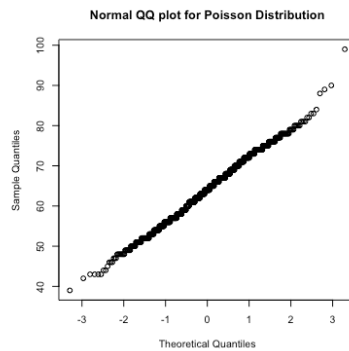
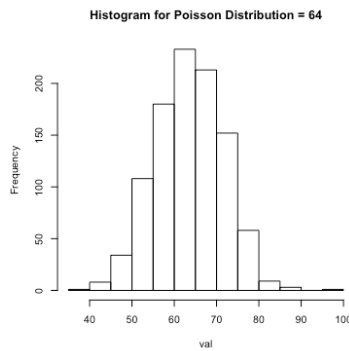
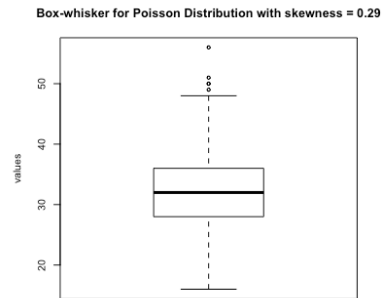
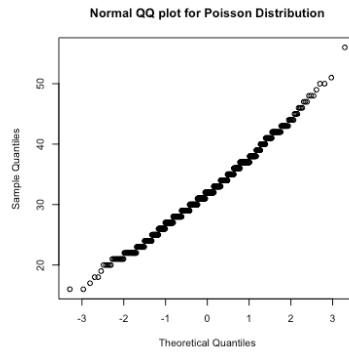
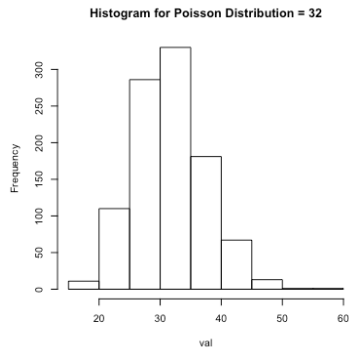
Remember that λ represents the mean of a discrete (counting) variable. At what size mean is Poisson data no longer skewed, relative to normally distributed data? You might run this 2 or 3 times, with different seeds; this number varies in my experience.

```
set.seed(54321)  
par(mfrow=c(1,3))  
lambda <- 2  
rpois <- {}  
  
while(lambda <= 64){  
  val <- rpois(1000, lambda = lambda)  
  skew <- round(skewness(val, na.rm = TRUE),2)  
  rpois[lambda-1] <- skewness(val)  
  
  # attaching all three plots together  
  par(mfrow=c(1,3))  
  
  # histogram  
  hist(val, main = paste0("Histogram for Poisson Distribution = ", lambda))  
  # QQ-norm  
  qqnorm(val, main = "Normal QQ plot for Poisson Distribution")
```

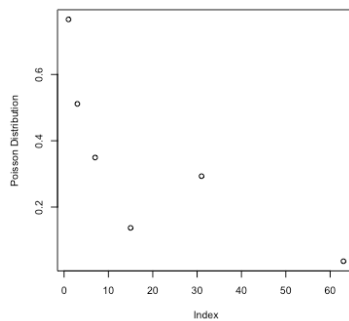
```
# box-whisker
```

```
boxplot(val, main = paste0("Box-whisker for Poisson Distribution with  
skewness = ", skew), ylab = "values")  
lambda <- lambda*2}
```





```
plot(rpois,ylab="Poisson Distribution")
```



If you do this in SAS, create a data table with data columns each representing a different μ . You can see combined histogram, box-whisker and QQ-norm, for all columns, by calling

```
proc univariate data=Distributions plot;
run;
```

At what μ is skewness of the Poisson distribution small enough to be considered normal?

Exercise 4

Part a

Write a function that accepts a vector `vec`, a vector of integers, a main axis label and an x axis label. This function should 1. iterate over each element i in the vector of integers 2. produce a histogram for `vec` setting the number of bins in the histogram to i 3. label main and x-axis with the specified parameters. 4. label the y-axis to read Frequency, `bins =` and the number of bins.

Hint: You can simplify this function by using the parameter `...` - see `?plot` or `?hist`

Part b

Test your function with the `hidalgo` data set (see below), using bin numbers 12, 36, and 60. You should be able to call your function with something like

```
hidalgo.data =
read.table("~/Desktop/work/GradSchool/Summer2020/STATS600/week9/hidalgo.dat",
header = T, sep = ",")
bin <- c(12,36,60)
hidalgo <- hidalgo.data[,1]
i=1

plot.histograms <- function(hidalgo, bin, main, xlab){
  par(mfrow=c(1,length(bin)))
  for(i in 1:length(bin)){
    print(hist(hidalgo, breaks = bin[i], main = main, xlab = xlab, ylab =
paste0('Frequency, bins = ', bin[i]))))}

plot.histograms(hidalgo,bin, main="1872 Hidalgo issue",xlab= "Thickness
(mm)")

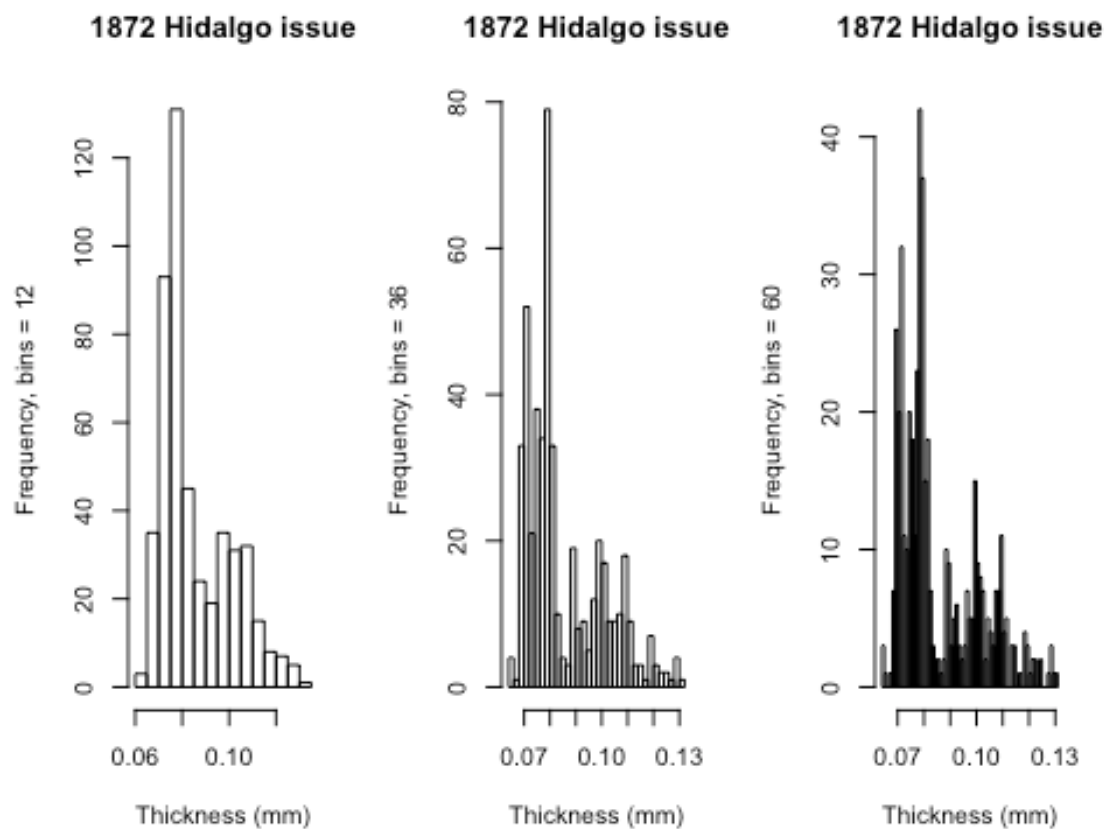
## $breaks
## [1] 0.060 0.065 0.070 0.075 0.080 0.085 0.090 0.095 0.100 0.105 0.110
## [12] 0.115 0.120 0.125 0.130 0.135
##
## $counts
## [1] 3 35 93 131 45 24 19 35 31 32 15 8 7 5 1
##
## $density
## [1] 1.2396694 14.4628099 38.4297521 54.1322314 18.5950413 9.9173554
## [7] 7.8512397 14.4628099 12.8099174 13.2231405 6.1983471 3.3057851
## [13] 2.8925620 2.0661157 0.4132231
##
## $mids
## [1] 0.0625 0.0675 0.0725 0.0775 0.0825 0.0875 0.0925 0.0975 0.1025 0.1075
## [11] 0.1125 0.1175 0.1225 0.1275 0.1325
##
## $xname
```

```

## [1] "hidalgo"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"

## $breaks
## [1] 0.064 0.066 0.068 0.070 0.072 0.074 0.076 0.078 0.080 0.082 0.084
## [12] 0.086 0.088 0.090 0.092 0.094 0.096 0.098 0.100 0.102 0.104 0.106
## [23] 0.108 0.110 0.112 0.114 0.116 0.118 0.120 0.122 0.124 0.126 0.128
## [34] 0.130 0.132
##
## $counts
## [1] 4 1 33 52 21 38 34 79 33 10 4 3 19 8 9 5 12 20 17 9 9 10 18
## [24] 9 3 3 1 7 3 2 2 1 4 1
##
## $density
## [1] 4.132231 1.033058 34.090909 53.719008 21.694215 39.256198 35.123967
## [8] 81.611570 34.090909 10.330579 4.132231 3.099174 19.628099 8.264463
## [15] 9.297521 5.165289 12.396694 20.661157 17.561983 9.297521 9.297521
## [22] 10.330579 18.595041 9.297521 3.099174 3.099174 1.033058 7.231405
## [29] 3.099174 2.066116 2.066116 1.033058 4.132231 1.033058
##
## $mids
## [1] 0.065 0.067 0.069 0.071 0.073 0.075 0.077 0.079 0.081 0.083 0.085
## [12] 0.087 0.089 0.091 0.093 0.095 0.097 0.099 0.101 0.103 0.105 0.107
## [23] 0.109 0.111 0.113 0.115 0.117 0.119 0.121 0.123 0.125 0.127 0.129
## [34] 0.131
##
## $xname
## [1] "hidalgo"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"

```

```
## $breaks
## [1] 0.064 0.065 0.066 0.067 0.068 0.069 0.070 0.071 0.072 0.073 0.074
## [12] 0.075 0.076 0.077 0.078 0.079 0.080 0.081 0.082 0.083 0.084 0.085
## [23] 0.086 0.087 0.088 0.089 0.090 0.091 0.092 0.093 0.094 0.095 0.096
## [34] 0.097 0.098 0.099 0.100 0.101 0.102 0.103 0.104 0.105 0.106 0.107
## [45] 0.108 0.109 0.110 0.111 0.112 0.113 0.114 0.115 0.116 0.117 0.118
## [56] 0.119 0.120 0.121 0.122 0.123 0.124 0.125 0.126 0.127 0.128 0.129
## [67] 0.130 0.131
##
## $counts
## [1] 3 1 0 1 7 26 20 32 11 10 20 18 11 23 42 37 15 18 7 3 2 2 1
## [24] 2 10 9 3 5 6 3 2 3 7 5 5 15 9 8 7 2 5 4 3 7 7 11
## [47] 4 5 0 3 3 0 1 0 4 3 1 2 2 0 2 0 0 1 3 1 1
##
## $density
## [1] 6.198347 2.066116 0.000000 2.066116 14.462810 53.719008 41.322314
## [8] 66.115702 22.727273 20.661157 41.322314 37.190083 22.727273 47.520661
## [15] 86.776860 76.446281 30.991736 37.190083 14.462810 6.198347 4.132231
## [22] 4.132231 2.066116 4.132231 20.661157 18.595041 6.198347 10.330579
## [29] 12.396694 6.198347 4.132231 6.198347 14.462810 10.330579 10.330579
## [36] 30.991736 18.595041 16.528926 14.462810 4.132231 10.330579 8.264463
## [43] 6.198347 14.462810 14.462810 22.727273 8.264463 10.330579 0.000000
## [50] 6.198347 6.198347 0.000000 2.066116 0.000000 8.264463 6.198347
```

```
## [57] 2.066116 4.132231 4.132231 0.000000 4.132231 0.000000 0.000000
## [64] 2.066116 6.198347 2.066116 2.066116
##
## $mids
## [1] 0.0645 0.0655 0.0665 0.0675 0.0685 0.0695 0.0705 0.0715 0.0725 0.0735
## [11] 0.0745 0.0755 0.0765 0.0775 0.0785 0.0795 0.0805 0.0815 0.0825 0.0835
## [21] 0.0845 0.0855 0.0865 0.0875 0.0885 0.0895 0.0905 0.0915 0.0925 0.0935
## [31] 0.0945 0.0955 0.0965 0.0975 0.0985 0.0995 0.1005 0.1015 0.1025 0.1035
## [41] 0.1045 0.1055 0.1065 0.1075 0.1085 0.1095 0.1105 0.1115 0.1125 0.1135
## [51] 0.1145 0.1155 0.1165 0.1175 0.1185 0.1195 0.1205 0.1215 0.1225 0.1235
## [61] 0.1245 0.1255 0.1265 0.1275 0.1285 0.1295 0.1305
##
## $xname
## [1] "hidalgo"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

to plot three different histograms of the hidalgo data set.

If you do this in SAS, write a macro that accepts a table name, a column name, a list of integers, a main axis label and an x axis label. This macro should scan over each element in the list of integers and produce a histogram for each integer value, setting the bin count to the element in the input list, and labeling main and x-axis with the specified parameters. You should label the y-axis to read Frequency, bins = and the number of bins.

Test your macro with the hidalgo data set (see below), using bin numbers 12, 36, and 60. You should be able to call your macro with something like

```
%plot_histograms(hidalgo, y, 12 36 60, main="1872 Hidalgo issue",
xlabel="Thickness (mm)");
```

to plot three different histograms of the hidalgo data set.

Hint: Assume 12 36 60 resolve to a single macro parameter and use %scan. Your macro definition can look something like

```
%macro plot_histograms(table_name, column_name, number_of_bins, main="Main",
xlabel="X Label")
```

Data

The hidalgo data set is in the file hidalgo.dat These data consist of paper thickness measurements of stamps from the 1872 Hidalgo issue of Mexico. This data set is commonly used to illustrate methods of determining the number of components in a mixture (in this case, different batches of paper). See <https://www.jstor.org/stable/2290118>, <https://books.google.com/books?id=1CuznRORa3EC&lpq=PA95&pg=PA94#v=onepage&q>

&f=false and

https://books.google.com/books?id=c2_fAox0DQoC&pg=PA180&lpg=PA180&f=false.

Some analysis suggest there are three different mixtures of paper used to produce the 1872 Hidalgo issue; other analysis suggest seven. Why do you think there might be disagreement about the number of mixtures?

##comments

I think it might be due to the paper thickness variations

Exercise 5.

We've been working with data from Wansink and Payne, Table 1:

Reproducing part of Wansink Table 1

Measure	1936	1946	1951	1963	1975	1997	2006
calories per recipe (SD)	2123.8 (1050.0)	2122.3 (1002.3)	2089.9 (1009.6)	2250.0 (1078.6)	2234.2 (1089.2)	2249.6 (1094.8)	3051.9 (1496.2)
calories per serving (SD)	268.1 (124.8)	271.1 (124.2)	280.9 (116.2)	294.7 (117.7)	285.6 (118.3)	288.6 (122.0)	384.4 (168.3)
servings per recipe (SD)	12.9 (13.3)	12.9 (13.3)	13.0 (14.5)	12.7 (14.6)	12.4 (14.3)	12.4 (14.3)	12.7 (13.0)

However, in Homework 2, we also considered the value given in the text

The resulting increase of 168.8 calories (from 268.1 calories ... to **436.9** calories ...) represents a 63.0% increase ... in calories per serving.

There is a discrepancy between two values reported for calories per serving, 2006. We will use graphs to attempt to determine which value is most consistent.

First, consider the relationship between Calories per Serving and Calories per Recipe:

Calories per Serving = Calories per Recipe / Servings per Recipe

Since Servings per Recipe is effectively constant over time (12.4-13.0), we can assume the relationship between Calories per Serving and Calories per Recipe is linear,

$$\text{Calories per Serving} = \beta_0 + \beta_1 \times \text{Calories per Recipe}$$

with Servings per Recipe = $1/\beta_1$

We will fit a linear model, with Calories per Recipe as the independent variable against two sets of values for Calories per Serving, such that

- Assumption 1. The value in the table (384.4) is correct.
- Assumption 2. The value in the text (436.9) is correct.

We use the data:

```
Assumptions.dat <- data.frame(  
  CaloriesPerRecipe = c(2123.8, 2122.3, 2089.9, 2250.0, 2234.2, 2249.6,  
3051.9),  
  Assumption1 = c(268.1, 271.1, 280.9, 294.7, 285.6, 288.6, 384.4),  
  Assumption2 = c(268.1, 271.1, 280.9, 294.7, 285.6, 288.6, 436.9))
```

and fit linear models

```
Assumption1.lm <- lm(Assumption1 ~ CaloriesPerRecipe, data=Assumptions.dat)  
Assumption2.lm <- lm(Assumption2 ~ CaloriesPerRecipe, data=Assumptions.dat)  
summary(Assumption1.lm)
```

```
##  
## Call:  
## lm(formula = Assumption1 ~ CaloriesPerRecipe, data = Assumptions.dat)  
##  
## Residuals:  
##      1      2      3      4      5      6      7  
## -7.0238 -3.8475  9.7610  4.7417 -2.5010 -1.3112  0.1808  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept)    25.477429   17.351550    1.468    0.202  
## CaloriesPerRecipe  0.117547    0.007466   15.745 1.88e-05 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 6.163 on 5 degrees of freedom  
## Multiple R-squared:  0.9802, Adjusted R-squared:  0.9763  
## F-statistic: 247.9 on 1 and 5 DF,  p-value: 1.879e-05
```

```
summary(Assumption2.lm)
```

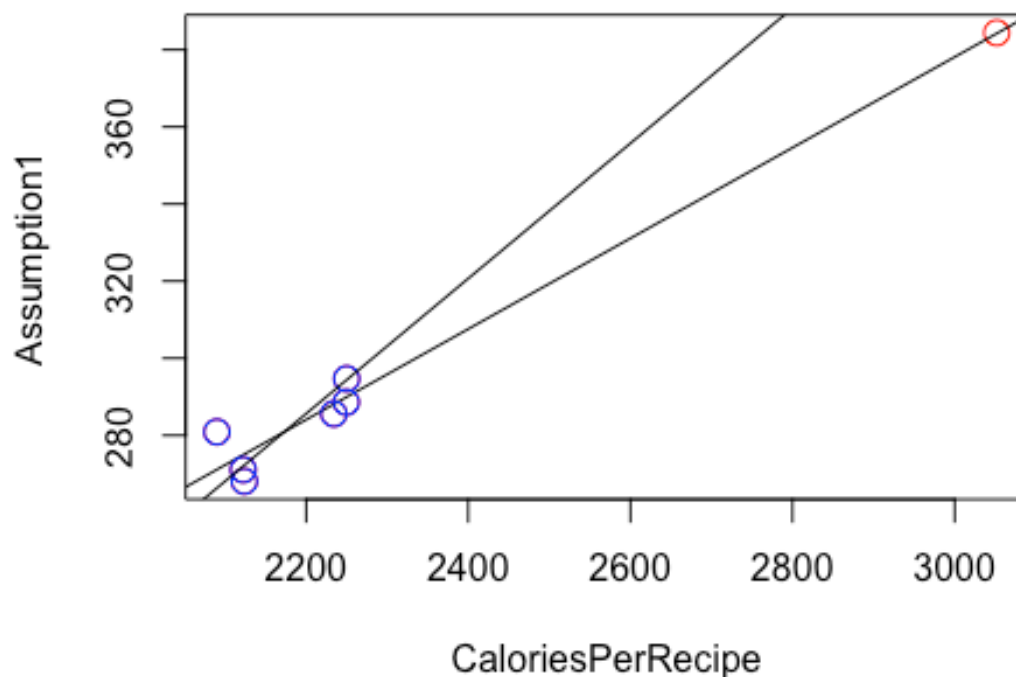
```
##  
## Call:  
## lm(formula = Assumption2 ~ CaloriesPerRecipe, data = Assumptions.dat)  
##  
## Residuals:  
##      1      2      3      4      5      6      7  
## -4.1798 -0.9169 14.5608  0.3051 -6.0261 -5.7248  1.9817  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      -99.891018  21.933161  -4.554  0.00609 **
## CaloriesPerRecipe  0.175238   0.009437  18.569  8.34e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.79 on 5 degrees of freedom
## Multiple R-squared:  0.9857, Adjusted R-squared:  0.9828
## F-statistic: 344.8 on 1 and 5 DF,  p-value: 8.336e-06
```

Part a.

Plot the regression. Use points to plot Assumption1 vs CaloriesPerRecipe, and Assumption2 vs CaloriesPerRecipe, on the same graph. Add lines (i.e. `abline`) to show the fit from the regression. Use different colors for the two assumptions. Which of the two lines appears to best explain the data?

```
attach(Assumptions.dat)
par(mfrow=c(1,1))
plot(Assumption1~CaloriesPerRecipe, cex = 1.5, type = 'p', col = 'red')
abline(Assumption1.lm)
points(Assumption2~CaloriesPerRecipe, cex = 1.5, col = 'blue')
abline(Assumption2.lm)
```

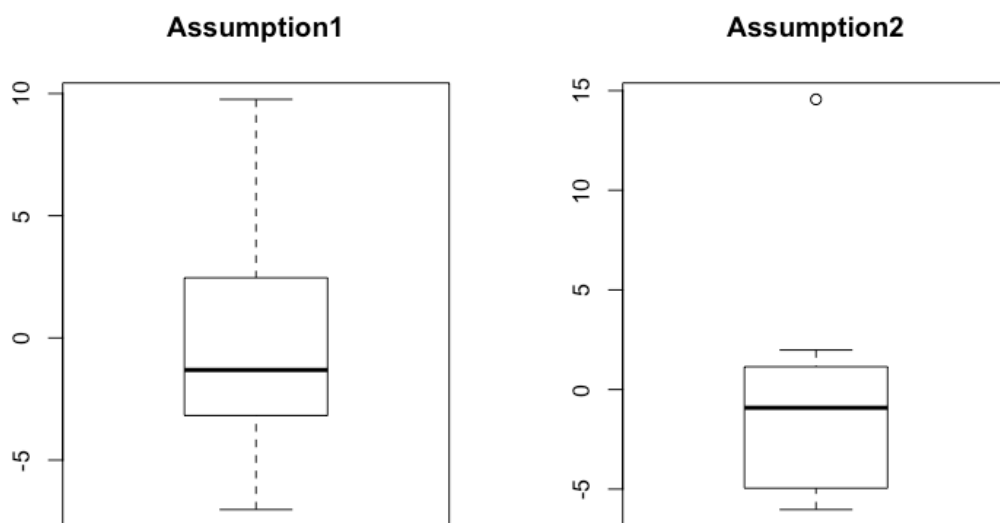


Part b.

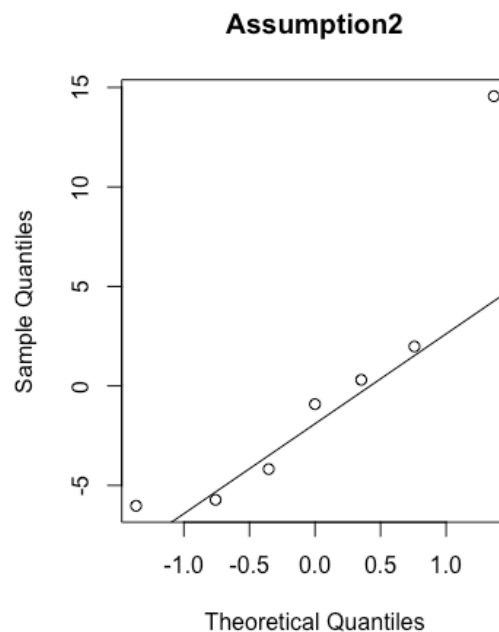
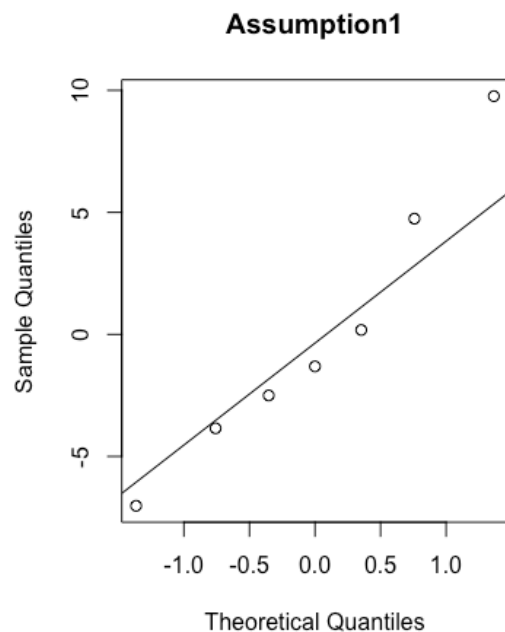
Produce diagnostic plots of the residuals from both linear models (in R, use `residuals(Assumption1.lm)`). qqnorm or box-whisker plots will probably be the most effective; there are too few points for a histogram.

Use the code below to place two plots, side by side. You can produce more than one pair of plots, if you wish.

```
#Box-whisker plots  
par(mfrow=c(1,2))  
boxplot(residuals(Assumption1.lm), main = "Assumption1" )  
boxplot(residuals(Assumption2.lm), main = "Assumption2" )
```



```
#QQnorm plots  
par(mfrow=c(1,2))  
qqnorm(residuals(Assumption1.lm), main = "Assumption1" )  
qqline(residuals(Assumption1.lm))  
qqnorm(residuals(Assumption2.lm), main = "Assumption2" )  
qqline(residuals(Assumption2.lm))
```



From these plots, which assumption is most likely correct? That is, which assumption produces a linear model that least violates assumptions of normality of the residual errors? Which assumption produces outliers in the residuals?

I've included similar data and linear models for SAS in the SAS template. If you choose SAS, you will need to modify the PROC GLM code to produce the appropriate diagnostic plots.
 ###comments Assumption 2 has quite a few anomalies which leads me to think assumption 1 is correct since its plot is within margin.