9 Additional Graphs Homework

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.

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(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.

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.

Part b

Plot the histograms for each distribution. Use `par(mfrow=c(2,2))` in your code chunk to combine the four histogram 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))
```

Part c

Repeat Part b, but with QQ-norm plots.

```
par(mfrow=c(2,2))
```

Part d

Repeat Part b, but with box-whisker plots.

```
par(mfrow=c(2,2))
```

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 `lambda`, from 2 to 64, doubling `lambda` each time. For each 'lambda' 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 `lambda` 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 `lambda` 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.

```
par(mfrow=c(1,3))
```

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

```
plot.histograms(hidalgo.dat[,1],c(12,36,60), main="1872 Hidalgo issue",xlab= "Thickness (mm)")
```

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&lpg=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?

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?

Part b.

Produce diagnostic plots 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 wise.

```
par(mfrow=c(1,2))
```

```
par(mfrow=c(1,2))
```

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.