STAT22000, Autumn 2013 Homework 1

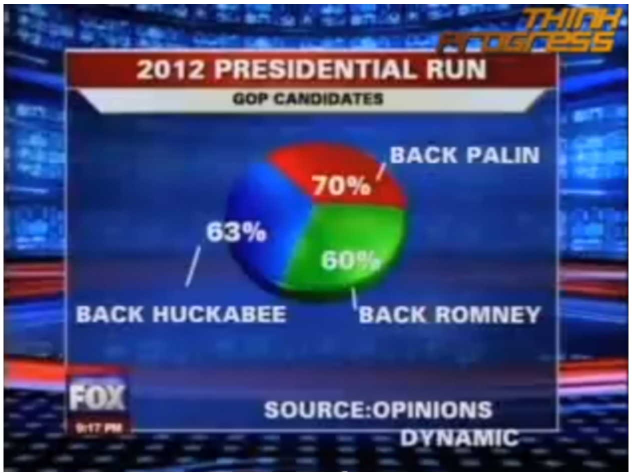
All page, section, and exercise numbers below are for the course text (Moore, McCabe and Craig, Introduction to the Practice of Statistics, 7th edition).

Reading: Section 1.1-1.2
Problems for Self-Study: (Do Not Turn In. Solutions are at the end of the textbook.)

1. Exercise 1.25 on p.23 (part (d) in particular)
2. Exercise 1.33 on p.25
3. Exercise 1.39 on p.26-27
4. Exercise 1.67 on p.46

Problems to Turn In: due Wednesday, Oct. 9, in class

1. (8 total points) The following pie chart comes from Fox News (Chicago) on November 23rd, 2009. (Here is the video `http://www.youtube.com/watch?v=-rbyhj8uTT8`)



   (a) (3 points) What is wrong with the pie chart?

   (b) (5 points) Make a correct graph to display the information on the Fox News pie chart.

2. (20 total points) The following is a stem-and-leaf plot of the cost of grocery purchases to the nearest dollars from a sample of 50 shoppers.

```
0 | 2
0 | 6689
1 | 0112234444
1 | 556788899
2 | 00113
2 | 6789
3 | 012333
3 | 679
4 | 03
4 | 5
5 | 12
5 |
6 | 24
6 | 9
```

   (a) Comment on the shape of the stemplot.
   (b) Which quartile (Q_1 or Q_3) do you expect to be farther from the median (Q_2)? Why?
   (c) Check your answer in (b) by computing the three quartiles.
   (d) Will the mean for these data be greater or small than the median? Why?
   (e) Use the 1.5 × IQR rule to check for outliers. How low would the highest amount need to be for it to be an outlier according to this rule?
   (f) Draw a modified boxplot by hand for this data set.

3. (22 total points) Daily rainfall in millimeters (mm) was recorded over a 47-year period in Turramurra, Sydney, Australia. For each year, the day with the greatest rainfall as identified.

The data are in the file rainfall.txt on Chalk. Use the R computing software to answer the following questions about the data.

(a) Find the mean, and the five-number summary of the rainfall values using the command summary().

(b) Do the numerical summaries in part (a) suggest a symmetric or skewed distribution for the greatest rainfall day in a year? Give two reasons for your conclusion.

(c) Explore the distribution of rainfall values by creating 3 histograms of the values, each with a different bin widths Use these plots to comment on the shape of the distribution and any values that you consider to be potential outliers.

(d) Are there any potential outliers among the rainfall values according to the 1.5 × IQR rule?

(e) Draw a boxplot of the rainfall values using the boxplot() function in R. Does this plot look as you expected given your answers above? As you explain, refer to at least two of items (a) to (d).

Some R Help for the Rainfall Problem. First follow the guideline on this webpage

http://www.stat.uchicago.edu/~yibi/R/Rtutorial.html

to change the working directory of R to the directory you store the file rainfall.txt. Then you can load the data into R using the command

```
> raindata = read.table("rainfall.txt", header=TRUE)
```

R will then store the data in a "data frame" called raindata. You see the content of raindata by

```
> raindata
```

in which you can see the data frame raindata contains only one variable: rainfall. In general, a data frame may contain several variables. The following are some instructions for part (a)(c)(e) of the problem.

(a) You can find the five-number summary of the data as well as the mean by the command

```
> summary(raindata)
```

(c) To draw histograms, use the command:

```
> hist(raindata$rainfall)
```

Here "raindata$rainfall" means the variable "rainfall" in the data frame "raindata". In general, data1$var1 means the variable "var1" under the data frame "data1".
Note if you type

```
> hist(raindata)
Error in hist.default(raindata) : 'x' must be numeric
```

you will get an error message because R only makes histograms for variables, not for data frames.
The command below doesn't work either because the variable rainfall is invisible to R because it is inside raindata

```
> hist(rainfall)
Error in hist(rainfall) : object 'rainfall' not found
```

R only knows the data frame **raindata**. One must tell R which data frame to look for the variable.
If you need to constantly refer to a data frame, one way to save some labor is to "attach" to the data frame.

```
> attach(raindata)
```

After attaching to it, all variables under the data frame "`raindata`" become visible to R. You can run the command below directly without typing "`raindata$`" again and again.

```
> hist(rainfall)
> hist(rainfall, breaks = 10)
> hist(rainfall, breaks = 5)
```

You may try different values of "breaks."
You can also specify the range of the histogram and the size of class intervals. For example, the command below creates a histogram covering the range 0 to 4000 and the size of class intervals is 250.

```
> hist(rainfall, breaks = seq(0,4000,by=250))
```

You can also specify the break points of class intervals explicitly, like

```
> hist(rainfall, breaks = c(400,800,1200,1600,2000,2400,2800,4000))
```

Be sure to include axes labels (including units) and titles for your histograms. Check the R help file using the command

```
> ?hist
```

to see how to change the axes labels and graph titles

(e) Read the R help file yourself.

```
> ?boxplot
```

When you finish with one data frame, be sure to "`detach()`" the current data frame

```
> detach(raindata)
```

before you "`attach`" to other data frames. Sometimes variables in different data frames have common names, R will get confused.