Assignment 5
More on Data frames, basic statistics and plotting
Wednesday Feb. 13, 2013
Due date: Fri. Feb. 24, 2013

During the course of the class, problems will become increasingly word-based with decreasing help on how to use R to accomplish your objectives. The hope is that over time, you will become comfortable transforming word prolblems (that do not include any information on R) into solutions derived using R.

In this assignment, you will continue "playing" with data.frames, perform some plotting operations on them and calculate some basic statistics.

Question 1: 30 pts
a) 10pts. Read the data set related to FSU football results from the file "fsu_football.RData" which can be found on the course website. This file contains data from 2007 to 2011, including variables for the year, the location (home vs. away), the opponent, outcome (win,lose), FSU score, opponent score, and game attendance. Make sure the headers in your data frame are the same as in the file by appropriately using the arguments to the function used to read the data. Create a variable that is FSU score minus opponent score (positive indicates a win, negative a loss) and add it as an extra column to your dataframe, using `cbind()`. Call this new column `result`. (If needed, use Google to find examples of how to use `cbind()` or try "?cbind" combined with `example(cbind` to see the examples from the help page being executed. Note that some examples are more sophisticated than needed in this exercises.)

b) 10pts. We will assess whether or not there is a home-field advantage and whether or not fan support, as measured by attendance, affects the outcome of the game. That is, is there a difference in mean result between home and away games? Compute the mean `result` for both home and away games. Which is larger? Create `result` a single plot with `result` against year that contains data for home games (in blue) and data for away games (in red). Do you discern any difference in the distribution of points for home and away games?

c) 10pts. Enhance the plot by adding meaningful labels, creating your axis labels and axis numbers with a 24 point font, and making sure that the font is of the `serif` family.

d) 10 pts. There is an alternate way of plotting the data above. Try

```
plot(fsu-opp ~ year, data=d)
```

The first argument is called a formula (we'll get to that later during class.) The second argument (data) is the name of a data frame. When the second argument is present, there is no need to write `d$fsu-d$opp` since R understands that the name of the dataframe is the value of the `data` argument.
Replace `year` by `factor(year)` and describe the resulting plot, called a box plot.

What are the characteristics of the box plot and what is displayed? Describe the plot you see. Are there any years substantially different from the others? If so, in what way? (these last two questions are not answered using R, but rather, by looking at the generated plots with a critical eye.

Question 2: 20 pts
Use the "attitude" data set within R, do the following.

a) 10pts. Add an additional factor column to the dataset with value "LOW" and "HIGH". "HIGH" refers to a row where the salary is higher than the mean salary raise. Name the new column "level". Save this new dataset to a comma-delimited file called "04_raise.RData". (Hint: you could use write.csv).

(b) 10pts. Show statistically (using the `t.test()` function) and graphically (using `plot()` whether the category of the amount of a raise received (low/high) significantly affects overall departmental rating. (Consider `t.test`.) Use the plot command using the form:

```
plot(vector ~ factor)
```

where `vector ~ factor` is called a formula (look it up, and describe what you find in your assignment report). What kind of plot is produced? Explain what you see and draw conclusions. Read about this kind of plot on the web if necessary. One way to do this is go to images.google.com, and use the terms "plot" and "R", and scroll until you see a plot similar to what you obtained. Then click on the image and read the document that contains the images.

Question 3: 30 pts

(a) 15 pts. Create a sample of size 500 from a uniform distribution that takes values in the range of -3 to 2 (we call this uniform distribution $U[-3, 2]$.) (Hint: these values are real numbers, not integers). Compute the mean value. Repeat this 5 times, and compute the standard deviation of these mean values. Repeat the experiment with a sample size of 5000. Compare the two standard deviations. Which one is larger?

(b) 15 pts. Starting from a uniform distribution $U[-1, 3]$ (sample values range from -1 to 3), plot the cumulative distribution function (CDF). (Hint. The CDF is also called the area under the histogram curve. It is calculated with the function `pnorm(x)`, which represents the area under the histogram $hist(y)$ for y smaller or equal to x. Look up one or two online tutorials to help you if needed.)

Create appropriate labels, make sure the symbols are the proper size. Overlay on this plot, the cumulative function of a uniform distribution $U[-2, 3]$. The first curve should be a blue solid line with width 3, the second curve should be red symbols.

Question 4: 20 pts Briefly answer the following (2 pts per question):

(a) What is one of the main differences between a data frame and a matrix

(b) What command can one use to add columns to a data frame?

(c) What command can one use to add rows to a data frame?

(d) What is the difference between the functions `read.csv` and `read.table`.

(e) What is the purpose of the `t.test` function?

(f) What is the purpose of the `t.shapiro` function?

(g) What does the function `apply` used for?

(h) What function is used to generate random integers?

(i) How does one display multiple plots on a single page?

(j) What does one display multiple lines on a single plot? (first give the command to display the first line, and then describe how to create the additional lines.)