Statistical Computing and Simulation

Spring 2017

Assignment 1, Due March 14/2017

1. (a) Use “*scan*” to read the data file ‘Draft Lottery’ from web page ***The Data and Story Library*** (http://lib.stat.cmu.edu/DASL/) that contains a mixture of character and numeric data. Using the “data.frame” to input the data from the story “[*Home Field Advantage*](http://lib.stat.cmu.edu/DASL/Stories/HomeFieldAdvantage.html)*.*” Then, output the data to a text file and an Excel (or csv) file.

(b) Change the missing values ‘¥’ to ‘NA.’ Then, draw a boxplot to describe the variables of home winning percentage and away winning percentage for the American League and for the National League.

(c) It is believed that the team with higher home winning percentage has better chance of winning the World Series. Check whether this result! (Hint: You can use the function “is.na” to remove the “NA” observations.)

(d) Draw scatterplots for the American League and the National League separately, using the home winning percentage as x-axis and the away winning percentage as y-axis. Also, use the function “identify” to mark unusual observations and give your comments on these observations.

2. (a) Use the commands “date” and “cat” to print the time you do this homework, e.g.,

*Today’s date is: Mon Mar 07 2017*

*The time now is: 09:13:00*

You need to write down the commands.

(b) Try the following commands in R and explain the outputs:

*(today <- Sys.Date())*

*format(today, "%d %b %Y") # with month as a word*

*(tenweeks <- seq(today, length.out=10, by="1 week")) # next ten weeks*

*weekdays(today)*

*months(tenweeks)*

*as.Date(.leap.seconds)*

1. (a) Plot the line graph of the following function on the interval [0, 6]:



(b) Use the command “abline” to plot the tangent lines at the minimum and 

1. Use the function “*rnorm*” to create 100 random numbers from normal distribution with mean and variance equal your birth month and birth date. Next, compute sample averages and variances of these observations, and averages of largest 40 observations. Continue this procedure 5,000 times and use the function “*apply*” to compute the sample averages and variances without using loop. Also, test if the averages of each set of 100 observations are equal to your birth month. (Bonus: Store the p-values of testing if they are from U(0,1)!)
2. (a) Plot *sin(x)* against x, using 200 values of x between π and −π, but do not plot any axes yet (use parameter *axes=F* in the call to plot.) Add a y-axis passing through the origin and horizontal labels. Add an x-axis with tick-marks from π to −π increments of π/4, twice the usual length.

(b) Similar to (a), but plot the functions *sin(x)* and *cos(x)* in the same plot, using the function *matplot* (i.e., matrix plot, or multiple plots). Also, use the function “legend” to explain the meaning of each line.

1. Use the command “readLine()” to input the text file on my website, as shown in the notes.
	1. Calculate the number of words and the number of different words in the file.
	2. Find the top ten two-word phrases and calculate their numbers of occurrences.
	3. Identify the locations of top three phrases and comment on their distribution. (Note: Check if they are normally or uniformly distributed.)