HU Extension **Assignment 02** E-185 Big Data Analytics

Handed out: 02/15/2013 Due by 17:25PM on Friday, 02/22/2013

Problem 1. Write an R function that would allow you to generate a matrix (data frame) with `m` columns and `n` rows and populate every column of that matrix with randomly generated values that fall between values `a` and `b`. Choose some arbitrary values for parameters `a` and `b`. Use your function to generate a matrix with 4 columns and a reasonably large value of the number of rows `n`, e.g. `1000-2000`. Plot histograms of data in two columns of your matrix and demonstrate to yourself that data are uniformly distributed, between selected values `a` and `b`. We are not performing any statistical tests for "uniformity" of the distribution. Just present data graphically in two different plots or histograms and take a look.

Problem 2. Try to find a way to present two distributions contained in any two of columns of your matrix on a single plot. To do that you might want to export the distribution data from two columns into two stand-alone vectors of equal length. Plot one distribution first using a call to `plot(x,y1)` function, where vector x contains the "predictor" or the parameter vector with values between `a` and `b` you selected above. To add the next curve (distribution `y2`) try invoking function `lines(x,y2)`. To improve your diagram, present two curves in different color and add labels on x and y axis, as well as the title to your graph. Try adding the distribution from the third column to your graph.

Problem 3. Write a simple R function that will determine the median of the distribution contained in a vector. Apply your function to a column of your matrix and compare the result with the value obtained with R provided function `median()`.

Problem 4. Using the function from **Problem 1**, generate a matrix with 50 columns. Add 5 new columns to your matrix. Let the first additional column contain the sum of the first 10 columns divided by 10. Let the second additional column contain the sum of the first 20 columns divided by 20, and so on until the fifth additional column contains a sum of all 50 original columns divided by 50. Present the distributions of data in those added columns on one plot. I hope that you will see how distributions gradually become less rugged and start approaching a bell curve.

Problem 5. For the distribution contained in the last (5^{th}) additional column use R supplied functions to determine the mean value and the standard deviation. Write a simple R function which would generate a Gaussian curve with provide mean and standard deviation. Plot that Gaussian function on the same plot with the distribution contained in that 5^{th} additional column. You might need to scale (multiply by a factor) the Gaussian in order to get a reasonable overlap of two distributions. Hopefully, you have demonstrated the Central Limit Theorem.

Problem 6. Plot the binomial distribution for p = 0.3, p = 0.5 and p = 0.7 and the total number of trials n = 60 as a function of k the number of successful trials. For each value

of p, determine 1st Quartile, median, mean, standard deviation and the 3rd Quartile. Present those values as a vertical box plot with the probability p on the horizontal axis.

SUBMISSION INSTRUCTIONS:

Your main submission should be an MS Word document containing your code, results produced by that code and brief textual descriptions of what you did and why. Typically, you just copy your code and results from the R console and past them into the Word document. Start with this text of homework assignment as the template. Please add any other files that you might have used or generated.

Package everything into an archive called E185_LastNameFirstNameHW02.zip. Naming your file properly is important. We download many files and if they are all named Assignment01.zip it becomes hard not to overwrite and lose them. Please do not use archiving tools which do not produce ZIP files.
If you are using a Mac, please make sure that your files are READABLE to users of Windows. You are welcome to save your work as a PDF file, but please, always submit a Word document, as well. Upload your ZIP archive to the course web site. Every assignment has its drop box. If you miss the deadline, please submit your solution into the 00_AnyHW_WayLate Drop Box. Those assignments will be graded as well. **We will chop 10% of your grade for every day you are late**. Your grade for every assignment will be entered as a comment next to your submission.

If you have issues with the formulation of the assignment or the software you are using, please FIRST go to the Discussion Forum on the class web site: http://isites.harvard.edu/icb/icb.do?keyword=k93720 and check whether someone else raised the same issue and whether the answer is already there. If not, raise the issues yourself. A person from the class or a member of the teaching stuff will respond.

If the issue is not address for a while, please send an inquiry to cscie185@fas.harvard.edu. The discussion forum is a very important tool. We all learn from the discussions on the forum.

If we respond to your inquiry to the class email address, PLEASE DO NOT RESPOND WITH A THANK YOU NOTE. This is not a joke. We will take 2% of your grade for that week's assignment for every "thank you note".

If you have issues with the class web site, please let us know right away. In the past, we experienced issues with the visibility of various folders, upload permissions and so on. We will try to resolve such issues as soon as we hear about them. For some issues we depend on the university support services and delays are possible.