**AC52012 Research Methods – Week 2 Lab**

**Basic Statistic Operations on Data Sets; Custom Functions**

In order to complete data set, you will need to understand the commands from the RTutorial file or to read the ‘An Introduction to R’ document that can be found at <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>. You also need to refer to the lecture notes for definitions of mean, median, standard deviation, percentiles, quartiles and outliers, as well as about histograms and boxplots.

1. Take the Nile data set from R, which contains measurements of the annual flow of the river Nile at Aswan in the period of a hundred years, from 1871 to 1970. The data is given in 108m3. Data set here is given as a vector of 100 elements.  
   1. Find the minimum and the maximum value of the set. Plot two histograms of this data – one that divides the data into 10 intervals (which will, therefore, have 10 bars) and the other that divides the data into 5 intervals (therefore, having 5 bars). Colour the bars of the histogram in blue.   
        
      **Hint**: Look at the hist command. Pay attention to the break and col parameters of it.
   2. Plot two versions of a boxplot for this data. In the first version, the boxplot should extend to the extreme values (minimum and maximum, this is the version of the boxplot mentioned in the lectures). In the second version, boxplot should only extend to the data values that are not outliers.   
        
      **Hint**: Use the boxplot command. Which of the two versions does the default version of this function, where you do not supply any parameter other than the dataset to plot, produce?
   3. Find the mean, median and standard deviation of this set.  
        
      **Hint:** Look at the mean, median and sd functions.
   4. Take random samples of the data of sizes 5, 10, 20 and 50. Create a data frame that will contain the mean, median and standard deviation of each of these samples. Measure how do these compare with the mean, median and standard deviation of the complete data set?  
        
      **Hint:** Look at the sample function for taking random samples from the data set. data.frame function can be used to create a data frame. To access individual columns of the data frame exFrame, use exFrame[“colName”] where colName is the name of the column.  
        
      **Hint 2**: One way of doing this is to create each of the four samples individually, then calculate its mean, median and standard deviation, put these into vectors and then create a data frame for it. A more efficient solution is to create a function, say getSample(n), that will for an argument n return a random sample of size n from Nile vector. Then, create a vector of sample sizes (containing elements 5, 10, 20 and 50) and apply a function getSample to each element of this vector (using sapply high-order function). This way we get a list of 4 samples. Using sapply again, apply mean, standard deviation and median built-in functions to the list of samples to obtain vectors of means, standard deviations and medians. Then use data.frame function to create a data frame from these.

* 1. Now, instead of taking random samples, take the samples which comprise of the first 5, 10, 20 and 50 elements of the data set and find the mean, median and standard deviation for them. How do these compare with the mean, median and standard deviation of the complete set?  
       
     **Hint:** Look at the index vectors and accessing parts of the vector. The rest is very similar to the example above.
  2. Find the 5-point summary of the data. Write a function that will, for a given point, check if the point is an outlier or not. Using this function, check if there are any outliers in the data set.   
       
     **Hint:** Function summary can be used to return the 5-point summary of a data set. Look at the IQR function for finding the interquartile range and quantile for computing percentiles and quartiles.
  3. Add a new value 100 to the end of the Nile dataset. Check if this value is an outlier.  
       
     **Hint:** Use function length to determine the length of Nile vector. Look at the tutorial file on how to add an element to the vector, or look at the append function.

1. The file EPL2021.csv contains information about all of the matches played in season 2021-22 in the English Premier League. It contains various pieces of data, such as number of goals, corners, shots (total and on targets), fouls and yellow and red cards. Of all this information, we are interested in total number of fouls and total number of corners. Home and away corners are stored in HC and AC columns, and home and away fouls in HF and AF columns.  
   1. Read this data into R data frame  
        
      **Hint:** Use read.csv function.
   2. Create two vectors from this data frame – corners, that will contain number of corners in each of the matches and fouls that will contain the number of fouls in each of the matches.  
        
      **Hint:** Look at the methods to extract columns from data frames. Extract HC and AC columns as vectors and add them together. Do the same with HF and AF
   3. Plot the histograms of corners and fouls.
   4. Find the mean, standard deviation, Q1, Q3 and IQR of corners and fouls.
   5. Find out if there are any outliers in these two data sets. Can you find out which of the teams were involved in the games that had numbers of fouls or corners that were the outliers?