STAT 40001/ MA59800 **Statistical Computing** **Fall 2017**

Homework 3

Name:

Due : October 5, 2017 PUID:

Instruction: Please submit your R code along with a brief write-up of the solutions (do not submit raw output containing ERRORs). Some of the questions below can be answered with very little or no programming. However, write code that outputs the final answer and does not require any additional paper calculations.

Q.N. 1) Generate 100 random numbers from normal distribution with mean 100 and standard deviation 10. How many observations are within one, two and there standard deviations from the mean? Compare your findings with the empirical rule.
According to the empirical rule **68%**, **95%** *and* **99.7%** *data reside within one, two and three standard deviation of the mean respectively. Does your data meet this rule?*

Q.N.2) The `abd` package in R contains data sets related to Biological studies. The data frame `TwoKids` in the `abd` package has of the information about the number of boys in two-child families. Display the information contain in the data choosing appropriate graphical method.

Q.N. 3) Air Quality Data Set for May 1973, from Chambers et al. (1983) consists of daily readings of air quality values from May 1, 1973 to September 30, 1973, but here are included only the values for May. Below are the variables listed in the data:

```
X1- Solar Radiation in Longleys  at Central Park
X2-  Average windspeed (in miles per hour)  at La Guardia Airport
X3-  Maximum daily temperature (in Fahrenheit) at La Guardia Airport
Y-   Mean ozone concentration (in ppb)  at Roosevelt Island
```

The data set is provide in the Blackboard as `airmay`
a) Import the data in R and remove all missing values
b) Calculate the numerical summary of each variables
c) Display the histograms for each variables in a **single plot** (Note: With the par( ) function, you can include the option mfrow=c(nrows, ncols) to creating desired number of spots for graphs)
d) Construct a 90% confidence interval for ozone concentration (in ppb) at Roosevelt Island.

Q.N.4) Patients with advanced cancers of the stomach, bronchus, colon, ovary or breast were treated with ascorbate. Data are available on the DASL web site *http://lib.stat.cmu.edu/DASL/Datafiles/CancerSurvival.html*
a) Import the Data set in R-readable format(You may first save and then import it using `read.table`)
b) Create side-by-side boxplots to compare the survival times with respect to the organ affected by the cancer.
d) Calculate the summary statistics of the survival times with respect to the organ affected by the cancer.

Q.N. 5) Results from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment conditions is provided in the data frame `PlantGrowth` in the R dataset.
a) How many observations are recorded in the data set?
b) What is the mean of each of the control and treatment conditions?
c) Test the hypothesis whether there is a significance difference between the treatment 1 and treatment 2.

Q.N. 6) The `Gasoline` data in `RSADBE` package include data frame with 25 observations on the 12 variables.
a) Access the data print the variables in the study
b) The variable `x11` is the Type of transmission (A-automatic, M-manual) and $\boldsymbol{y}$ is the mileage. Test the hypothesis whether the manual cars have higher milage.

Q.N. 7) The babies data frame in the `UsingR` packages has a collection of variables taken for each new mother in a Child and Health Development Study. The variable age contains the moms age and the variable `dage` contains the dads age for several babies. Do a significance test of the null hypothesis of equal ages against a one-sided alternative that dads are older

Q.N. 8) A person makes a doctor appointment, receives all the instructions and doesn't show up for appointment, Who to blame? Data set containing some information including the age, gender are provided in the data set (*Noshow*). (Other variable names are self-explanatory).
a) Import the data in R and identify its dimension
b) Print the variables included in the dataset.
c) Display the Age distribution by gender creating parallel box plot.
d) Test whether female are more likely to miss the appointment than male.
e) Test whether the SMS reminder helped not to miss the appointment.
f) Are female older than male? Perform the test.
(Hint: xtabs function in R will be useful to create Cross-Tabulation)