STAT 40001/ MA 59800 **Statistical Computing** **Fall 2017**

Homework 4

Due : October 31, 2017

Name:
PUID:

Instruction: Please submit your R code along with a brief write-up of the solutions. 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) Access the data from url *http://www.stat.berkeley.edu/users/statlabs/data/babies.data* and store the information in an object named **BABIES** using the function *read.table()*
a) Create a `CLEAN` data set that removes subjects if any observations on the subject are "unknown". Note that if the values are unknown `bwt, gestation, parity, height, weight` and `smoke` are quoted as 999, 999, 9, 99, 999, and 9 respectively. Store the modified data set in an object named `CLEAN`.
b) Create side-by-side boxplots to compare the birth weights of babies for both smoking and non-smoking mothers.
c) Calculate the five number summaries of the birth weights of babies of both smoking and non-smoking mothers.

Q.N. 2) A college bookstore claims that, on average, a college student will pay **$100** per class for textbooks. A student group investigates this claim by randomly selecting ten courses from the course catalog and finding the textbook cost for each course. The data collected is

$$\mathbf{140, 125, 150, 102, 143, 170, 120, 94, 53, 115}$$

a) At **0.05** level of significance is there an enough evidence to prove that the test book cost is greater than **$100** per class.
b) Construct a **95%** confidence interval for the average test book cost per course.
c) Construct a **90%** confidence interval for the average test book cost per course.

Q.N. 3) Water-quality researchers wish to measure biomass/chlorophyll ratio for phytoplankton(in milligrams per liter of water). There are two possible test, one less expensive then the other. To see whether the two tests give the same results, ten water sample were taken and each was measured both ways. Table below provide the measurements. Perform a test to see if there is a difference in the means of the measured amounts. Please list all the assumptions you made to perform the test.

Method 1	45.9	47.6	54.9	38.7	35.7	39.2	45.9	43.2	45.4	54.8
Method 2	48.2	64.2	56.8	47.2	43.7	45.7	53.0	52.0	45.1	57.5

Q.N. 4) Data set "quine" from MASS package children from an Australian town is classified by ethnic background, gender, age, learning status and the number of days absent from school. The columns "Eth" indicates whether the student is Aboriginal or Not ("A" or "N"), and the column Sex indicates Male or Female ("M" or "F").
a) Print first five observations of the data.
b) Is the proportion of aboriginal female different from that of male?

Q.N. 5) The dataset *HairEyeColor* in R contains classifications of students by gender, color and eye color.
a) How many students are included in the data set?
b) Display the information using mosaicplot(HairEyeColor, col=c(1,2))
b) Is hair color independent of eye color for men?
c) Is hair color independent of eye color for women?

Q.N. 6) The data frame *TestScores* in the PASWR packages gives the test grades of 20 students taking a basic statistics course.
(a) Use the function EDA() on the data. Can normality be assumed?
(b) Perform the test for normal distribution.

Q.N. 7) The on-base percentage (OBP) indicates how successful a baseball player as a batter. The data set OBP in UsingR contains the OBP for the year 2000. Test whether these data come from a normal distribution.