**STAT 40001/ STAT 59800 Statistical Computing Fall 2020**

**Test 1**

**Name:**

**PUID:**

*This exam consists of 6 questions. Please answer only five questions (Q.N. 1 is required but you may choose any four questions from Q.N. 2 to Q.N. 6). Please provide the R codes that have been used to perform the calculations and graphics along with the interpretation of the output. It is important that you distinguish between the R code and the description. You may use different fonts or different colors.*

**Q.N. 1)** **Short answer questions**

1. Generate 200 random numbers from a normal distribution with mean 5 and standard deviation 6. Please print only first 5 observations and last 5 observations.

**b)** For 97 countries in the world, data are given for birth rates, death rates, infant death rates, life expectancies for males and females, and Gross National Product are provided in the link below

<http://ww2.amstat.org/publications/jse/datasets/poverty.dat.txt>

Missing values are indicated by \*. Please clean the dataset by removing all \*s. What is the dimension of the Clean data?

**c)** Graduate student enrollments in Statistics and Biostatistics departments in 2009 are provided in StatisticsPhD data set in *Lock5withR* package. The list does not include combined departments of mathematics and statistics and does not include departments that did not reply to the AMS survey. Please access the data and display the number of full-time graduate students based on the department (Statistics Vs. Biostatistics).

**d)** The primes dataset in the UsingR package contains set of prime numbers in [1, 2003]

(i) How many prime numbers are in [1, 2003]?

(ii) How many prime numbers are in [1,100]?

(iii) How many prime numbers are in [100, 1000]?

**e)** Daily rainfall (in millimetres) was recorded over a 47-year period in Turramurra, Sydney, Australia. For each year, the wettest day was identified (that having the greatest rainfall). The data are provided in the link below

<http://www.statsci.org/data/oz/sydrain.txt>

Draw a histogram of the rainfall. Please make sure to change color, insert the title, labels etc.

**Q.N. 2)** The breast cancer database obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg are stored as biopsy in *MASS* package in R.

a) Access the data and identify its dimension.

b) The missing values are are marked as “NA”. Please remove all NAs and create a dataset

and name it *Clean .*

c) How many observations are there in the *Clean* data set?

d) The variable class classify the status of the disease as "benign" or "malignant". How many cases are benign and how many are malignant?

e) Display the class variable graphically.

**Q.N. 3)** The dataset survey in MASS package contains the responses from a sample of students at the University of Adelaide to a number of questions.

**a)** Import the data in R and determine how many questions the survey includes.

**b)** The variable sex refers to the gender of the student ("Male" and "Female") and the variable Smoke indicates how much the student smokes. ("Heavy", "Regul" (regularly), "Occas" (occasionally), "Never".) Create a frequency table to determine number of male and female students in each of the four levels of Smoke.

**c)** Test whether the male students are older than female.

**d)** Test whether there is a difference in pulse rate based on the gender.

**Q.N. 4)** WHAT'S WHAT AMONG AMERICAN COLLEGES AND UNIVERSITIES?

This is the subject of the 1995 Data Analysis Exposition sponsored by the Statistical Graphics Section of the American Statistical Association. The AAUP data includes average salary, overall compensation, and number of faculty broken down by full, associate, and assistant professor

ranks. The data are provide in the link <http://jse.amstat.org/datasets/aaup.dat.txt>

Here are some of the first few variables

1 - 5 FICE (Federal ID number)

7 - 37 College name

38 - 39 State (postal code)

40 - 43 Type (I, IIA, or IIB)

44 - 48 Average salary - full professors

49 - 52 Average salary - associate professors

53 - 56 Average salary - assistant professors

1. Import the data in R and determine its dimension.
2. The Missing values are marked as “\*”. How many observations have at least one missing value?
3. Display the average salary of full professor, associate professor and assistant professor using parallel box plot.
4. How many universities from the state of Indiana (IN) are included in this study?

**Q.N. 5)** The dataset concerning hepatitis are provided in the link below

<https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/hepatitis.data>

In this dataset the second column represents the age and third column gender (1- male and 2-female).

1. Import the data in R and identify its dimension.
2. Are there any missing values? If so remove them.
3. Test the hypothesis whether there is an age difference between male and female.
4. Construct a 90% confidence interval for the age difference between male and female.

**Q.N. 6)** Data from a study comparing brain size and intelligence is available in the attached file (Brain) with this test in the Brightspace.

a) Import the Data set in R-readable forma and print first 5 observations.

c) It appears that there are few missing values marked as “¥”. How many observations have at least one missing value?

d) Display the MRI counts based on the gender by choosing appropriate graph

e) Do we have enough evidence to conclude that female’s MRI count is less than male’s MRI count?