**INSTRUCTIONS**

**• For the entire Problem Set 2, you need to save and submit your work in Canvas as a single .R file.**

**• Important: Use comments throughout your code to:**

**- Indicate which question you are answering**

**Example: #Question 1A, #Question 1B, etc**

**- Annotate your work**

**Examples: #Calculate the mean for ‘height’**

**#P(being on time) = 1/6**

**#Apply binomial distribution**

**\*\*\*If you do not provide annotation, we will not be able to give you full credit for solving the problems.**

**• If you have any questions about what is being asked or what you need to do in order to solve the problems, please reach out to a TA or instructor as soon as possible. If it is necessary to provide additional information or corrections, updated information will be posted on the PS1 Assignment page in Canvas.**

**30pts**

1. For Problem 1 questions, use the dataset for “Risk Factors Associated with Low Infant Birth Weight.” You should use R to examine the dataset and answer the following questions.

Additional information about the dataset:

• Source: Hosmer, D.W. and Lemeshow, S. (1989) \_Applied Logistic Regression.\_ New York: Wiley

• This data frame contains the following columns:

‘low’ indicator of birth weight less than 2.5 kg.

‘age’ mother's age in years.

‘lwt’ mother's weight in pounds at last menstrual period.

‘race’ mother's race (‘1’ = white, ‘2’ = black, ‘3’ = other).

‘smoke’ smoking status during pregnancy. 1 = yes, 0 = no

‘ptl’ number of previous premature labours.

‘ht’ history of hypertension.

‘ui’ presence of uterine irritability.

‘ftv’ number of physician visits during the first trimester.

‘bwt’ birth weight in grams.

5 pts

A. How many observations are in the dataset?

5 pts

B. Examine the features, determine what type of variable each represents, and indicate whether each one is discrete or continuous. Then, go on to determine the distribution or descriptive statistics as appropriate:

For Discrete Variables:

• Indicate whether the feature is nominal, ordinal, or binary

• How many levels each variable has

For Continuous variables:

• Determine the mean, standard deviation, and median

5 pts

C. How many individuals older than 30 smoke?

5 pts

D. Plot a histogram for birth weight.

5 pts

G. Calculate the probability of randomly selecting an individual that has either a low birth weight or a mother who was a smoker.

5 pts

H. Calculate the probability of randomly selecting an individual that is white and has more than 3 physician visits during the first trimester.

**30 pts**

2. Use R to answer the following probability questions.

10 pts

A. What is the probability that given a positive mammogram exam, a woman has a positive cancer diagnosis? Assume that the breast cancer incidence rate is 1%, the positivity rate for the exam if a patient has cancer is 90%, and there is a false positive rate of 8% for the exam.

10 pts

B. For every attempt to call your friend, there is a 70% probability of actually speaking with them. Calculate the probability of having exactly 12 successes in 20 attempts.

10 pts

C. If a sample of test scores is normally distributed with a mean of 42 and a standard deviation of 8, what percent of the scores is:

(i) Greater than 25?

(ii) Smaller than 31?

(iii) Between 25 and 31?

**20pts**

3. Markov Chain:

physical exercise training method A is used only 5% of the time, a person using method A will stay with this method 85% of the time, and a person not using method A will switch to method A about 65% time. At the beginning of the experiment only 5% of people used method A.

1. Generate a transition matrix for this Markov chain
2. Generate a transition plot (using R or by hand as an image it’s valid)
3. Plot the change in the probabilities over time for both methods until the 10th time unit.

**20 pts**

4. Random Walk:

Another simpler example of a random walk is a one-dimensional random walk. first we place a marker at zero (our initial state), we flip a coin, if it lands on heads, the marker is moved one unit to the right (1), if it lands on tails it is moved one unit to the left.

#### 1. Generate a function that randomly draws from our initial state and populates a vector with the different transitions.

#### 2. Generate a plot that shows 500 independent one-dimensional walks, differentiating walks that end above 0 or below 0.

#### 3. What is the frequency of walks that ended in a positive cumulative count, in zero, or negative?