**MATLS 3J03: STATISTICAL METHODS IN BIOMEDICAL ENGINEERING**

**ASSIGNMENT #1**

**Topics:**

Read and Explore Data

Data Cleaning

Data Visualization

Descriptive Statistics

Probability

Discrete Probability Distributions

**Total marks: 100**

**Submission Instructions:** Electronic submission to the IBEHS 4C03 Avenue to Learn Assignment 1 folder.

**Due:**

**Assignment Submission Format:**

File Names: StudentLastNameFirstNameAssignment#.doc/pdf/py//ipynb etc.

Both your name and student numbers should appear at the top of the document. A single document in pdf or word should be submitted for the graded version of your submission. Submit fully supported answers to the questions in the single document including the plots you made in python, and refer to coding that can be embedded or appended to the document. The python file used to generate the plots and any provided answers should be submitted with any assignment as a pdf or word document conversion of the python file and added at the end of the document or could be used to answer questions directly through using fully annotated or Markdown generated answers. Supplemental documents can be submitted in the supplemental dropbox.

**Problem 1: [25 points/100 points] Descriptive Statistics**

**Hospital dataset**: The data in the Hospital.csv dataset is collected on people discharged from a hospital as part of a retrospective chart review of antibiotic usage. Use the Hospital dataset to answer the following problems.

1. It is of clinical interest to know if the duration of hospitalization stay (in days of stay) is affected by whether a patient has received antibiotics (Antibiotics=1, No Antibiotics=0). Upload the Hospital dataset. Find the mean, median, 5-number summary information, sample variance, and sample standard deviation of the two Antibiotic group and the No Antibiotic Group. Discuss and show all of the steps to finding these statistics. Discuss the distributions of the duration of stay.
2. Next, explore this question descriptively using boxplots for each group (use the closest point to 1.5\*IQR for the whisker length). Discuss each of the elements of the boxplot. What does the boxplot tell you about the central tendency and the variability of the data? Are there any outliers present in the two subsamples? What impact do they have on the data? Discuss how this affects the answer to question a) if you remove any identified outliers.

**Problem 2: [25 points/100 points] Data Exploration and Data Cleaning**

**Heart Disease dataset:** Import the heart\_disease\_needs\_data\_cleaning dataset into your statistical software. This dataset is based on the Cleveland Heart Disease dataset, where 303 patients had clinical attributes recorded and the presence or absence of coronary artery (heart) disease (Heart\_Disease (0 is no heart disease and 1 is heart disease. This diagnosis is based on angiographic disease status with Value 0 < 50% diameter narrowing and Value 1 ≥ 50% diameter narrowing).

The variables that we will focus on is

* Age (in years old)
* Slope (the slope of the peak exercise ST segment of the ECG: Value 1 = upsloping; Value 2 = flat; Value 3 = down-sloping

Review the dataset and inspect the data for data cleaning/data preprocessing.

1. Let’s first look for completeness. For each column of data of interest, how many data values and data rows contain missing data? Briefly explain what you could do with these data values. Now go ahead and delete the rows that have any missing values.
2. Next, look for data-type validity. Since we care about Age and Slope, let’s check that the data type matches what you would want the data to be used for in an analysis of heart disease. What are the data types for Age and Slope in the dataset? If the data type does not match what you want it to be, discuss what the data type is that you want and why, and the reasons that you find in the dataset for any datatype mismatch, and correct this. (Hint: There are a lot of way to do this: If you need to remove any values that do not match your desired data type, you can delete the row of data based on the datatype mismatch once you identify, or you can change mismatches to NaNs and then deal with those missing values by deleting by row). In the end, you want the data types to be what you want specified.
3. Now look at data validity. What do you think the limits of plausible data would be for Age and Slope? Go ahead and clean the data based on deleting any nonsense values. Again, delete entire rows to end up with a clean dataset of complete data on Age and Slope.
4. Choose a graph to represent Age, Slope, or Age and Slope and why you chose the visualization that you did. Show and discuss your graph(s).

**Problem 3: [25 points/100 points]: Probability**

1. A circuit is shown in the following figure. The switches for the circuit operate independently of one another, and the probability of each switch is:

* The probability of A, B being closed is 0.75, 0.9 respectively.
* The probability of C being open is 0.20.

What is the probability that the circuit works? (Note: work these questions out by hand and show your work.)

Shape

Description automatically generated with medium confidence

1. From the following General Probability Rules, discuss the ones that you used to answer Question 3.1 above in the context of the question.
   1. The range and sum of probabilities in a distribution
   2. The Complement Rule
   3. The Multiplication Rule for independent events and for dependent (or conditional) events
   4. The Addition Rule for mutually exclusive events
   5. The Addition Rule for non-mutually exclusive events
   6. The Conditional Probability Rule: Bayes Rule

**Problem 4: [25 points/100 points] Discrete Probability Distributions**

1. Wastewater samples contain a high level of microplastic particles in 10% of tests. High levels were defined as >0.10g/L. Let X = number of samples high in micro-plastic particles in the next 18 samples tested. Assume samples are independent of each other with regards to particulate concentration.
2. Discuss if the Binomial distribution is a reasonable model to answer questions about sample particulate concentrations and why. Make sure to Identify the random experiment and the random variable and state any assumptions that you make.
3. Specify the parameters of the binomial distribution and discuss the shape of the distribution in this example.
4. What is the probability that all the samples have high levels of particulates?
5. What is the probability that 2 of the samples have high levels of particulates?
6. What is the probability that at least 4 of the samples have high levels of particulates?
7. There are crack-initiating defects along a steel pipe. Your company wants to model the number of defects in a length of pipe knowing that there is a mean of 2.3 defects per meter. Assume that the number of defects in each interval is independent of other intervals.
8. Discuss if the Poisson distribution is a reasonable model to answer questions about defect counts along the pipe and why? Make sure to Identify the random experiment and the random variable and state any assumptions that you make.
9. Specify the parameters of the Poisson distribution and discuss the shape of the distribution in this example.
10. What is the probability that there are 2 defects in 1 meter of pipe?
11. What is the probability that there are 10 defects in 5 meters of pipe?
12. What is the probability that there is at least 1 defect in 2 meters of pipe?