PH227: AI and Data Science
Midsem Exam Data Science Part: Autumn 2025

Time/Date: 13:30 – 15:30 PM, Sep 13, 2025
Venue: LC 102 + LT 102+LT 103
Maximum Marks: 30

Instructions: You must bring your own fully charged laptop using which you have to write computer codes in Python for the problems given in the question paper and then upload the corresponding .ipnyb or .py files on the submission portal on Moodle before leaving the exam venue. You are free to use: (a) internet, (b) chatGPT or any other AI utility, and (c) any Python library of your choice. However, you cannot communicate with each other during the exam, nor can you submit someone else's solutions as your own. To prevent this, we will apply plagiarism check on all the submitted files.
IMPORTANT: This question paper has two parts: you have to attempt 2 problems from part I, and all 4 from part II. Thus, you have to attempt six (6) problems in total.

Part I (Python Programming): Attempt any two problems; each problem is worth 5 marks.

1. Write a Python program to print out all the integers between 1 and N which are divisible by 3 in the reverse order. That is the largest number should be printed first, and smallest the last. Note that $N > 0$ is a user-specified integer.

2. Write a Python program to generate a 2-dimensional 100×100 array with elements (x_1, x_2), where both x_1 and x_2 are random numbers of normal distributions. Assume that for x_1, $\mu_1 = 2.0$ and $\sigma_1 = 1.0$, and for x_2, $\mu_2 = -1.0$ and $\sigma_2 = 0.5$. The program should also print out a scatter plot corresponding to (x_1, x_2).

3. Evaluating the value of π using random numbers. Implement the following algorithm in a Python program:

   (a) Generate N pair of uniform random numbers (x, y), in the range $0 \leq x \leq 1,\ 0 \leq y \leq 1$, where $N(> 0)$ is a user specified large integer. Each (x, y) pair will be called a point.

   (b) Figure out how many of these points lie on or inside the unit circle centered at the origin. That is, how many of these points satisfy the inequality $x^2 + y^2 \leq 1$. Let this number be N_1.

   (c) Rules of probability tell us that for $N \to \infty$, $N_1/N \to \pi/4$. Thus you should get an estimate for π by computing $4N_1/N$. Note that with increasing N, this ratio will approach the true value of π.

Part II: Data Science + Statistics: All problems of this part are compulsory.

1. You are provided with a CSV file named synthetic_dataset.csv, which contains 315 rows and the following 5 columns: (a) ID – Unique identifier for each entry, (b) Category, a categorical variable with values like A, B, C, D, (c) Value1, An integer in the range (10-100), (d) Value2, A floating-point column with values in the range (1.0–50.0), and (e) Status, a categorical variable with values Active or Inactive. The dataset has been deliberately corrupted with: (a) 10 missing values spread across the columns (excluding ID), and (b) 15 duplicated rows. Write a Python program which will

   (a) Load the CSV file into a pandas DataFrame.

   (b) Display the shape of the dataset (rows and columns).

   (c) Find and print the total number of missing values in each column.

   (d) Display all rows that contain at least one missing value.

   (e) Check for duplicate rows and print the total number of duplicates.

   (f) Remove duplicate rows from the DataFrame.

   (g) Save the cleaned DataFrame into a new file named synthetic_dataset_cleaned.csv. (6 marks)

2. Write a Python program which will:

   (a) Generate 10,000 random samples of size $n = 30$ from an exponential distribution with parameter $\lambda = 1.0$. For this purpose, use the function numpy.random.exponential.

   (b) Compute the sample mean for each of the 10,000 samples.

   (c) Plot a histogram of the 10,000 sample means.

   (d) Superimpose on this histogram a normal distribution curve with the mean equal to the population mean ($\mu = 1/\lambda = 1.0$) and the standard deviation equal to:

   $$\sigma_x = \frac{\sigma}{\sqrt{n}}.$$

   For the present case, the population standard deviation is $\sigma = 1.0$.

   (e) Repeat the experiment with sample sizes $n = 5$, 30, and 100. Compare the histograms and discuss how the distribution of the sample means changes with increasing sample size. (7 marks)

3. Write a Python script to perform the following statistical analysis on the data given in the file student_scores.csv, which has 4 columns and 800 rows:

   (a) Ask the user whether to use the data of column TestScoreA or TestScoreB

(b) Once the user chooses the column, calculate and print the values of μ, σ, skewness, and kurtosis for the data contained in that column, and also make a histogram plot of the data. Calculate these quantities from using their definitions, by using the for loops. If you wish, you can also verify compare your results with those computed using appropriate functions in numpy and scipy. However, your marks will be determined by your own code mentioned above. (3 marks)

4. Use the csv file of the previous problem. Write a Python program which will

   (a) Ask the user whether to use the data of column TestScoreA or TestScoreB

   (b) ask the user to whether to perform a z-test or a t-test

   (c) ask the user how many sample points (n) to use from the 800 data points of the chosen column. Randomly draw n points from the chosen column to form a sample of n data points.

   (d) ask the user the value of the significance level α

   (e) calculate and print the value of the chosen statistic (z or t), along with the corresponding p value.

   (f) write your conclusion either "reject H0" or "fail to reject H0", based on a two-tailed test. H_0 is that the calculated population mean (of 800 points of the chosen column) is correct (4 Marks)