Plaksha University

DS3001 - Advanced Statistics (Jan-May 2026)

Group Assignment - 1

Instructions:

- Group size should not be more than **FOUR.** Inform the names of the members of your group to TAs before submission.
- Submit the assignments with codes
- **Include the prompts you have used for AI to get the answer (if any AI is used).**
- **Due Date: February 24, 2026. Time: 23.59**

1. Suppose $X \sim \mathcal{N}(0,1)$ and $Y = g(X) = X + 0.3X^2$.
   (a) Discuss whether the CDF of Y admits a closed-form expression. If yes, derive the CDF of Y. If not, propose and implement a practical computational approach to evaluate the CDF of Y.
   (b) Verify your proposed computational approach results using the Probability Integral Transformation (PIT).
2. Let $X \sim Geometric(p),\ 0 < p < 1$ and $Y \sim Exponential(\lambda),\ \lambda > 0$ random variables.

(a) Simulate large ($n = 1000$) independent random samples from both the distributions (by assuming some specific values for p and λ) and empirically verify the lack-of-memory property by choosing some specific values for $m,\ n, s$ and t.

$$P(X > m + n \mid X > m) = P(X > n);$$

$$P(Y > s + t \mid Y > s) = P(Y > t)$$

(b) Quantify deviations from exact equality due to finite-sample effects.

(c) Use simulated conditional distributions to test the lack-of-memory property?

(d) Explore whether finite mixtures of geometric or finite mixtures of exponential distributions retain the lack-of-memory property. (you may consider mixture of two components)

3. (a) Explain the concepts associated with a Probability-Probability (P-P) plot and a Quantile-Quantile (Q-Q) plot used for graphically verifying the appropriateness of a distribution for the given data.

(b) Randomly generate 10, 100 and 1000 (three sets) observations from a $\mathcal{N}(100, 16^2)$ distribution and construct a normal Q-Q plot for each set of observations. Comment on the clustering of points along the line $y = x$.

(c) What is the slope of the line approximating these points?

(d) Randomly generate 1000 observations from the (i) Exponential distribution($\lambda = 2.5$) and (ii) Uniform distribution over (0,1). Construct the normal quantile plot in each case. Explain how they reveal the non-normality of the data in both cases.

(e) For case (ii) in (d), show that the Q-Q plot with uniform quantile is approximately linear. Construct the plot.

4. Generate a random sample of size 50 from two different normal distributions $\mathcal{N}_4(\boldsymbol{\mu}_1, \Sigma_1)$, and, $\mathcal{N}_4(\boldsymbol{\mu}_2, \Sigma_2)$, with

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 2 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 2 \\ -1 \\ 1 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 2.0 & 0.3 & 0.2 & 0.1 \\ 0.3 & 1.5 & 0.4 & 0.2 \\ 0.2 & 0.4 & 1.8 & 0.3 \\ 0.1 & 0.2 & 0.3 & 1.2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 5.0 & -1.8 & 0.0 & 0.0 \\ -1.8 & 0.9 & 0.0 & 0.0 \\ 0.0 & 0.0 & 3.5 & 1.6 \\ 0.0 & 0.0 & 1.6 & 0.8 \end{pmatrix}.$$

Mix these observations and try to classify them into one of the populations using Mahlanobis distance and Euclidean distance. Discuss the misclassification errors in each case. Briefly explain the method used for evaluating the misclassification error.

5. Use *Iris Data* (150 observations, 3 species, 4 continuous features; sepal length, sepal width, petal length, petal width) to carry out the clustering using Mahalanobis distance and Euclidean distance. Compare the efficiency of both the clustering approaches..