Kasarani Campus
Off Thika Road
P. O. Box 49274, 00101
NAIROBI
Westlands Campus
Pamstech House
Woodvale Grove
Tel. 4442212
Fax: 4444175

KIRIRI WOMENS' UNIVERSITY OF SCIENCE AND TECHNOLOGY
UNIVERSITY EXAMINATIONS, 2024/2025 ACADEMIC YEAR
FIRST YEAR, SECOND SEMESTER EXAMINATION
FOR MASTER OF SCIENCE IN APPLIED STATISTICS AND DATA ANALYTICS
KMA 5108: APPLIED MULTIVARIATE DATA ANALYSIS

DATE: 31ST JANUARY, 2025
TIME: 1:00 AM – 4:00 PM

INSTRUCTIONS TO CANDIDATES
ANSWER QUESTION ONE (COMPULSORY) AND ANY OTHER THREE QUESTIONS

QUESTION ONE: COMPULSORY (40 MARKS)

a) Let $\mathbf{X} = (X_1, X_2)'$ be a bivariate normal vector with known population covariance matrix

$$\Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$$

The sample mean vector based on a sample of size $n = 25$ is $\underline{\bar{X}} = (3,\ 5)'$

At 5% significance level, test the hypothesis $H_0 : \underline{\mu} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$ against $H_1 : \underline{\mu} \neq \begin{pmatrix} 2 \\ 4 \end{pmatrix}$

(5 Marks)

b) Consider a trivariate normal vector $\mathbf{X} = (X_1, X_2, X_3)'$ with mean vector $\underline{\mu}' = (1, -1, 2)'$ and $\Sigma = \begin{bmatrix} 7 & 3 & 2 \\ 3 & 4 & 1 \\ 2 & 1 & 2 \end{bmatrix}$.

Partition $\underline{X}$ as $X_1 = X_1$ and $\underline{X}_2 = (X_2, X_3)'$. Find the conditional density of X_1 given $\underline{X}_2$. Hence or otherwise, obtain the regression coefficients for the regression line of X_1 on X_2 *and* X_3 **(6 Marks)**

c) Given a trivariate normal density with variance-covariance matrix

$$\Sigma = \frac{1}{6}\begin{bmatrix} 4 & 2 & 1 \\ 2 & 7 & 1/2 \\ 1 & 1/2 & 7/4 \end{bmatrix}$$

Compute $R^2_{1.23}$ **(4 Marks)**

d) Let $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, suppose $\underline{X}$ is partitioned such that $\underline{X} = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

Explain how Canonical Correlation Analysis can be used to test the hypothesis $H_0 : \Sigma_{12} = 0$ against $H_1 : \Sigma_{12} \neq 0$ **(7 Marks)**

e) Explain the main objective in discriminant analysis. Hence or otherwise, derive the classification rule for classifying a new observation $\underline{X}$ into either P_1 or P_2 **(6 Marks)**

f) Suppose the data below is a random sample from a bivariate normal distribution

$$\underline{X}^T = \begin{pmatrix} 3 & 4 & 5 & 4 \\ 6 & 4 & 7 & 7 \end{pmatrix}^T$$

Obtain:

i. The maximum likelihood estimate of the mean vector **(3 Marks)**

ii. The maximum likelihood estimate of the covariance matrix **(4 Marks)**

g) A random sample of size 10 was obtained from a bivariate normal population with mean vector $\underline{\mu}$ and variance-covariance matrix Σ_0 (known) where $\Sigma_0 = \begin{bmatrix} 4 & 4.2 \\ 4.2 & 9 \end{bmatrix}$, $\underline{\mu}_0 = (6 \quad 5)'$. Given that $\overline{\underline{X}} = (5.8, \ 5.2)'$, carry out the test at $\alpha = 0.01$ level of significance for $H_0 : \underline{\mu} = \underline{\mu}_0$, $H_1 : \underline{\mu} \neq \underline{\mu}_0$ **(5 Marks)**

QUESTION TWO: (20 MARKS)

a) In a physiological research on a given flower from a plant, the following four measurements (in cm.) were taken on each flower from 100 plants; sepal length (X_1), sepal width (X_2), petal length (X_3) and petal width (X_4). The results of the experiment were as follows:

$$\overline{\underline{X}} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \overline{X}_4 \end{bmatrix} = \begin{bmatrix} 5.936 \\ 2.770 \\ 4.260 \\ 1.326 \end{bmatrix} \quad \text{and} \quad S = \begin{bmatrix} 0.260 & 0.085 & 0.183 & 0.054 \\ 0.085 & 0.098 & 0.083 & 0.091 \\ 0.183 & 0.083 & 0.221 & 0.073 \\ 0.054 & 0.091 & 0.073 & 0.039 \end{bmatrix}$$ with characteristic roots

$$\underline{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \overline{\lambda}_4 \end{bmatrix} = \begin{bmatrix} 0.4879 \\ 0.0724 \\ 0.0548 \\ 0.0098 \end{bmatrix}$$ and corresponding Eigen vectors

$$\beta = \begin{bmatrix} \underline{\beta}_1 & \underline{\beta}_2 & \underline{\beta}_3 & \underline{\beta}_4 \\ 0.68 & -0.66 & -0.26 & 0.10 \\ 0.31 & 0.56 & 0.72 & -0.22 \\ 0.62 & 0.34 & 0.62 & -0.31 \\ 0.22 & 0.33 & 0.06 & 0.91 \end{bmatrix}$$

i. Using empirical test, obtain the expression for the principal components **(5 Marks)**

ii. Also, obtain the percentage contribution of each component **(3 Marks)**

b) Let $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n$ be ***n*** independent observation vectors from a multivariate normal population with mean vector $\underline{\mu}$ and covariance matrix Σ. Define the sample mean vector $\overline{\underline{X}} = \frac{1}{n}\sum_{r=1}^{n} \underline{X}_r$ and sample covariance matrix as $S = \left(\left(S_{ij}\right)\right)$,

$$S_{ij} = \frac{1}{n}\sum_{r=1}^{n}\left(X_{ri} - \overline{X}_i\right)\left(X_{rj} - \overline{X}_j\right)$$

i. Derive the distribution of $\overline{\underline{X}}$ (Hint: show that $\overline{\underline{X}} \sim N\left(\underline{\mu}, \Sigma/n\right)$ **(5 Marks)**

ii. Show that the sample covariance matrix S is biased for Σ **(5 Marks)**

iii. Hence or otherwise, obtain the unbiased estimator for Σ **(2 Marks)**

QUESTION THREE: (20 MARKS)

a) The data below refers to nutritional contents of three diets. Variables measured are

$$\underline{\mathbf{Y}} = \left[\mathbf{Y}_1\left(Ascorbic\right), \mathbf{Y}_2\left(Riboflavin\right)\right]'$$

Diet A ($n_1 = 2$)	Diet B ($n_2 = 3$)	Diet C ($n_3 = 3$)
$\begin{bmatrix}0.25\\1.50\end{bmatrix}, \begin{bmatrix}0.59\\1.78\end{bmatrix}$	$\begin{bmatrix}1.85\\2.90\end{bmatrix}, \begin{bmatrix}3.50\\4.00\end{bmatrix}, \begin{bmatrix}1.80\\3.15\end{bmatrix}$	$\begin{bmatrix}0.74\\0.95\end{bmatrix}, \begin{bmatrix}1.25\\1.80\end{bmatrix}, \begin{bmatrix}0.95\\1.55\end{bmatrix}$

i. Write down appropriate statistical model for analyzing this data **(3 Marks)**

ii. Find the between groups (B) and within groups (W), SS and CP Matrices **(8 Marks)**

iii. Form a MANOVA Table **(5 Marks)**

iv. Test for equality of diet content at 0.1 level of significance **(4 Marks)**

QUESTION FOUR: (20 MARKS)

a) Two bivariate normal populations are mixed together. It was later decided that the two populations be separated. The parameters of the two distributions are

$P_1 : \underline{X} \sim N_2\left(\underline{\mu}_1, \Sigma\right)$ and $\underline{\mu}_1 = \left(6.2,\ 3.8\right)'$

$P_2 : \underline{X} \sim N_2\left(\underline{\mu}_2, \Sigma\right)$ and $\underline{\mu}_2 = \left(5.8,\ 3.5\right)'$ and $\Sigma = \begin{bmatrix}25 & 16\\16 & 16\end{bmatrix}$

Construct the optimal linear discriminant rule and classify a new observation $\underline{X} = \left(6.0,\ 3.4\right)'$ **(6 Marks)**

b) The table below shows laboratory results of three characteristics of soil chemical contents (measured in milliequivalents per 100 g) from some 10 different locations. The variables are

Y_1= available soil calcium,

Y_2 = exchangeable soil calcium,

Y_3 = turnip green calcium.

Location	1	2	3	4	5	6	7	8	9	10
$\mathbf{Y_1}$	35	35	40	10	6	20	35	35	35	30
$\mathbf{Y_2}$	3.5	4.9	30.0	2.8	2.7	2.8	4.6	10.9	8.0	1.6
$\mathbf{Y_3}$	2.80	2.70	4.38	3.21	2.73	2.81	2.88	2.90	3.28	3.20

Assuming normal distribution, use the provided data to obtain the following:

i. Mean vector **(3 Marks)**

ii. Dispersion matrix **(5 Marks)**

iii. Conditional distribution of Y_3 given $(Y_1, Y_2) = (27.0, 7.0)'$ **(6 Marks)**

QUESTION FIVE: (20 MARKS)

a) Consider two bivariate normal populations $\mathbf{X}_1 \sim N_2(\underline{\mu}_1, \Sigma)$ and $\mathbf{X}_2 \sim N_2(\underline{\mu}_2, \Sigma)$ where the covariance matrix Σ is known and the same for both populations. From the two populations, we collect two independent random samples of size n_1 and n_2 with sample means $\overline{\mathbf{X}}_1$ and $\overline{\mathbf{X}}_2$. You are tasked with testing the hypothesis: $H_0 : \underline{\mu}_1 = \underline{\mu}_2$ against $H_1 : \underline{\mu}_1 \neq \underline{\mu}_2$

i. Derive the test statistic for the hypothesis and explain how you would perform the test and how you would determine the rejection region. **(3 Marks)**

ii. Use the following data to perform the test at 5% significance level of α = 0.05:

$Sample\ 1:\ \overline{\mathbf{X}}_1 = \begin{pmatrix} 2.5 \\ 3.0 \end{pmatrix},\quad n_1 = 15$ and $Sample\ 2:\ \overline{\mathbf{X}}_2 = \begin{pmatrix} 2.0 \\ 2.8 \end{pmatrix},\quad n_1 = 20$ with a known covariance matrix $\Sigma = \begin{bmatrix} 1.5 & 0.6 \\ 0.6 & 1.2 \end{bmatrix}$ **(5 Marks)**

b) The dataset below contains measurements of two features (height and weight) from two different groups: Group 1 and Group 2.

Group 1		Group 2	
Height	**Weight**	**Height**	**Weight**
1.50	65	1.70	75
1.55	62	1.72	78
1.58	63	1.75	80
1.62	68	1.80	85

Your task is to perform a linear discriminant analysis (LDA) by answering the following questions:

i. Obtain the within-group covariance matrix. **(7 Marks)**

ii. Calculate the linear discriminant function coefficients. **(5 Marks)**