JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF BIOLOGICAL, PHYSICAL, MATHEMATICS AND ACTUARIAL SCIENCES
UNIVERSITY EXAMINATION FOR DEGREE OF BACHELOR OF SCIENCE IN ACTUARIAL SCIENCE
4th YEAR 2nd SEMESTER 2023/2024 ACADEMIC YEAR
MAIN REGULAR

COURSE CODE: WAB2412

COURSE TITLE: MULTIVARIATE METHODS

EXAM VENUE: **STREAM: (BSc. Actuarial Science)**

DATE: **EXAM SESSION: Jan-April 2024**
TIME: 2.00 HOURS

Instructions:

i. Answer questions one and any other two.
ii. Candidates are advised not to write on the question paper.
iii. Candidates must hand in their answer booklets to the invigilator while in the examination room.
iv. Where necessary, computations and data analysis to be done with R software.

QUESTION ONE (30 Marks)

a) Using the MILK TRANSPORT COST DATA given in QUESTION 5 (Data 5), test for multivariate normality of each group (gasoline trucks and diesel trucks as groups). How many outliers can you identify in each group using a 3D plot or a bivariate box plot?

(7 marks)

b) Let x_1, x_2, x_3, x_4 be five random variables with $\underline{\bar{x}}$ = [39.88 45.08 48.11 49.95] as the mean vector of a sample with n=150.

$$R=\begin{bmatrix} 1.0 & 0.7501 & 0.6329 & 0.6363 \\ & 1.0 & 0.6925 & 0.7386 \\ & & 1.0 & 0.6625 \\ & & & 1.0 \end{bmatrix}$$

(8 marks)

R is the correlation matrix. Perform principal component analysis using R and:

i. Give the eigenvalues and corresponding eigenvectors
ii. Write down all the sample cantered principal components
iii. What proportion of variance is explained by the first principal component?
iv. How many components would you select to explain the variation and why?

c) Using multivariate analysis of variance, analyse the following data for the concentration of three amino acids in centipede hemolymph (mg/100ml), asking whether the mean concentration of each is the same in males and female: (using the any three; Wilk's, pillar's trace, Lawley-Hoteling trace, Roy's maximum root, or Hoteling T^2)

(8 marks)

Male			Female		
alanine	aspartic acid	tyrosine	alanine	aspartic acid	tyrosine
7	17	19.7	7.3	17.4	22.5
7.3	17.2	20.3	7.7	19.8	24.9
8	19.3	22.6	8.2	20.2	26.1
8.1	19.8	23.7	8.3	22.6	27.5
7.9	18.4	22	6.4	23.4	28.1
6.4	15.1	18.1	7.1	21.3	25.8
6.6	15.9	18.7	6.4	22.1	26.9
8	18.2	21.5	8.6	18.8	25.5

d) Let random variables $\underline{x}^{!} = [x_1, x_2, x_3]$ be distributed as N_3 $(\underline{\mu}, \sum)$, where

$$\underline{\mu} = \begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix} \quad \sum = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{pmatrix}$$

i. Which variables are independent
ii. Find the correlation matrix of $\underline{x}$

iii. Find the distribution of z=$4x_1 - 6x_2 + x_3$
iv. Find the distribution of z=$\begin{pmatrix} x_1 - x_2 + x_3 \\ 2x_1 + x_2 - x_3 \end{pmatrix}$

(7 marks)

QUESTION TWO (20 Marks)

a) Let y be a random vector $\underline{y}^{!} = [y_1, y_2\ y_3]$ with mean vector and covariance matrix

$$\underline{\mu} = \begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix} \quad \sum = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 3 \\ 0 & 3 & 10 \end{pmatrix}$$

Given $\underline{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \quad \sum = \begin{pmatrix} y_1 + y_2 + y_3 \\ 3y_1 + y_2 - 2y_3 \end{pmatrix} and\ \underline{W} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 2y_1 - y_2 + y_3 \\ y_1 + 2y_2 - 3y_3 \\ y_1 + y_2 + 2y_3 \end{pmatrix}$

Find the following

i. Expectation: E$[\underline{z}]$ and E$[\underline{w}]$
ii. Covariance: cov$[\underline{z}]$ and cov E$[\underline{w}]$
iii. Covariance: cov E$[\underline{z}, \underline{w}]$

(8 marks)

b) A zoologist collected lizards and measured mass (in grams) and the snout-vent length SVL (in millimeters). The data for the lizards from the two genera, Cnemidophorous (C)- 20 observations and Sceloporus (S) -40 observations are below. Using the logarithmic (ln) transformations of the observations;

i. Find the mean vector and sample dispersion matrix for each genera
ii. By conducting univariate student's t-test to test for no difference in the individual means between the two genera for each variable, compare the result with the multivariate test on equality of mean vectors. (In other words demonstrate the efficacy of multivariate tests relative to their univariate counterparts).

(12 marks)

LIZARD DATA FOR TWO GENERA

C		S		S	
Mass	**SLV**	**Mass**	**SLV**	Mass	**SVL**
7.512	74	13.911	77	14.666	80
5.032	69.5	5.236	62	4.79	62
45.867	72	37.331	108	5.02	61.5
11.088	80	41.781	115	5.22	62
2.419	56	3.995	106	5.69	64
13.61	94	3.962	56	6.763	63
18.247	95	4.367	60.5	9.977	71
16.832	99.5	3.048	52	8.831	69.5
15.91	97	4.838	60	9.493	67.5
17.035	90.5	6.525	64	7.811	66
16.526	91	22.61	96	6.685	64.5
4.53	67	13.342	79.5	11.98	79
7.23	75	4.109	55.5	16.52	84
5.2	69.5	12.369	75	13.63	81
13.45	91.5	7.12	64.5	13.7	82.5
14.08	91	21.077	87.5	10.35	74
14.665	90	42.989	109	7.9	68.5
6.092	73	27.201	96	9.103	70
5.264	69.5	38.901	111	13.216	77.5
16.0902	94	19.747	84.5	9.787	70

QUESTION THREE (20 Marks)

a) Consider data matrix for n=3 for a bivariate distribution

$$X = \begin{bmatrix} 6 & 10 & 8 \\ 9 & 6 & 3 \end{bmatrix}$$

$$\bar{X} = \begin{bmatrix} 8 \\ 6 \end{bmatrix}$$

Evaluate the observed T^2 for $\underline{\mu}'_0 = [9 \quad 5]$. What is the sampling distribution of T^2 in this case? (10 marks)

b) Consider the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$$

And the derived correlation matrix

$$\rho = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 100 \end{bmatrix}$$

Determine the principal components for $\sum$ providing percentage of explained variability for each variate. (10 marks)

QUESTION FOUR (20 Marks)

a) The following are data on the percentage effectiveness of a pain reliever and the amounts of three different medications (in milligrams) present in each capsule;

medication A x_1	medication B x_2	medication AC x_3	percentage effectiveness y
15	20	10	47
15	20	20	54
15	30	10	58
15	30	20	66
30	20	10	59
30	20	20	67
30	30	10	71
30	30	20	83
45	20	10	72
45	20	20	82

45	30	10	85
45	30	20	94

Assuming the regression is linear; find the multiple regression models for response variable Y.

Is regression significant? Explain (8 marks)

b) Let x_1, x_2, x_3, x_4, x_5 be five random variables with $\underline{\bar{x}} = [.0054, .0048, .0057, .0063, .0037]$, n=100,

$$R = \begin{bmatrix} 1.00 & 0.557 & 0.509 & 0.387 & 0.462 \\ & 1.00 & 0.599 & 0.389 & 0.322 \\ & & 1.00 & 0.436 & 0.426 \\ & & & 1.00 & 0.523 \\ & & & & 1.00 \end{bmatrix}$$

(12 marks)

R is the correlation matrix. Perform principal component analysis using R and:

i. Give the eigenvalues and corresponding eigenvectors
ii. How many principal components explain more than 70% of the variation
iii. Obtain the principal component solution of the orthogonal factor model for two factors
iv. Obtain the maximum likelihood estimate for the solution of the orthogonal factor model suing two factors, how do they compare to the factor loading of the principal components solution in (ii).

QUESTION FIVE (20 Marks)

a) The following are the cholesterol contents in milligrams per package that four laboratories obtained for 6-ounces packages of three very similar diet foods:

	Diet food A	Diet food B	Diet food C
Laboratory 1	3.4	2.6	2.8
Laboratory 2	3.0	2.7	3.1
Laboratory 3	3.3	3.0	3.4
Laboratory 4	3.5	3.1	3.7

i. Considering diet foods as the only factor, test at α=0.05 level of significance whether the difference among the three-sample means can be attributed r=to chance
ii. Perform a two-way analysis of variance and test the null hypothesis concerning the diet foods and the laboratories at 0.05 level of significance

(8 marks)

b) In the first phase of a study of the cost of transporting milk from farms to diary plants, a survey was taken of firms engaged in milk transportation. Cost data on X_1 = fuel, X_2=repair and X_3= capital, all measured on a cost per mile basis are presented in the table below for n_1=36, gasoline and n_2=23 diesel truck.

i. Test for differences in mean cost vectors. Set α=0.01

ii. Construct 99% simultaneous confidence intervals for the pairs of mean components. Which cost, if any, appear to be quite different?

iii. Comment on the validity on the assumptions used in your analysis. Observations 9 and 21 for gasoline tricks have been identified as multivariate outliers. Repeat part i) with these observations deleted. Comment on the results. (12 marks)

MILK TRANSPORTATION COST DATA

GASOLINE TRUCKS			**DIESEL TRUCKS**		
X_1	X_2	X_3	X_1	X_2	X_3
16.4	12.43	11.23	8.5	12.26	9.11
7.19	2.7	3.92	7.42	5.13	17.15
9.92	1.35	9.75	10.28	3.32	11.23
4.24	5.78	7.78	10.16	14.72	5.99
11.2	5.05	10.67	12.79	4.17	29.28
14.25	5.78	9.88	9.6	12.72	11
13.5	10.98	10.6	6.47	8.89	19
13.32	14.27	9.45	11.35	9.95	14.53
29.11	15.09	3.28	9.15	2.94	13.68
12.68	7.61	10.23	9.7	5.06	20.84
7.51	5.8	8.13	9.77	17.86	35.18
9.9	3.63	9.13	11.61	11.75	17
10.25	5.07	10.17	9.09	13.25	20.66
11.11	6.15	7.61	8.53	10.14	17.45
12.17	14.26	14.39	8.29	6.22	16.38
10.24	2.59	6.09	15.9	12.9	19.09

10.18	6.05	12.14
8.88	2.7	12.23
12.34	7.73	11.68
8.51	14.02	12.01
26.16	17.44	16.89
12.95	8.24	7.18
16.93	13.37	17.59
14.7	10.78	14.58
10.32	5.16	17
8.98	4.49	4.26
9.7	11.59	6.83
12.72	8.63	5.59
9.49	2.16	6.23
8.22	7.95	6.72
13.7	11.22	4.91
8.21	9.85	8.17
15.86	11.42	13.06
9.18	9.18	9.49
12.49	4.67	11.94
17.32	6.86	4.44

11.94	5.69	14.77
9.54	16.77	22.66
10.43	17.65	10.66
10.87	21.52	28.47
7.13	13.22	19.44
11.88	12.18	21.2
12.03	9.22	23.09