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 2021/2022 ACADEMIC YEAR
MAIN REGULAR

COURSE CODE: WAB2412

COURSE TITLE: MULTIVARIATE METHODS

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

DATE: **EXAM SESSION: May-August 2022**
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)

Use the KEarrest data provided
(https://drive.google.com/drive/folders/1CZAS5ZgaTLAZufBxg4xBT8I4zspIpiv0?usp=sharing) to answer the following questions. The data set contains a hypothetical/simulated statistic, in arrests per 100000 residents for assault, murder, and rape in each of the Kenyan 47 counties in the year 2018. Also given is the percent of the population living in urban areas.

a) Plot a bivariate boxplot for the KEarrest. Explain the graph. (2 marks)
b) Obtain a correlation matrix making sense of the coefficients. (2 marks)
c) Compute the variance-covariance matrix (2 marks)
d) Evaluate the trace of the var-cov matrix (2 marks)
e) Compute the cross product matrix of the var-cov matrix (2 marks)
f) Find the determinant and inverse of the var-cov matrix (2 marks)
g) Obtain the eigen values and corresponding eigen vectors of the var-cov matrix (2 marks)
h) Provide spectral decomposition of the var-cov matrix (2 marks)
i) When is a matrix said to be symmetric positive definite? (2 marks)
j) Evaluate whether the var-cov matrix is symmetric positive definite (2 marks)
k) Convert the var-cov matrix to a correlation matrix, calculate the eigen values and eigen vectors and verify that the eigen vectors are orthogonal (2 marks)
l) Assume that X is the KEarrest data. Obtain $X^T X$. Use R to obtain the resulting matrix. (4 marks)
m) Which of the following are orthogonal to each other?

$$x = \begin{pmatrix} 1 \\ -2 \\ 3 \\ -4 \end{pmatrix} \quad y = \begin{pmatrix} 6 \\ 7 \\ 1 \\ -2 \end{pmatrix} \quad z = \begin{pmatrix} 5 \\ -4 \\ 5 \\ 7 \end{pmatrix}$$

(4 marks)

QUESTION TWO (20 Marks)

Use the dataset also available (here:
https://drive.google.com/drive/folders/1CZAS5ZgaTLAZufBxg4xBT8I4zspIpiv0?usp=sharing) to answer the following questions. The data frame contains the frequency of execution of 13 house tasks in the couple. This data/table is also available in ade4 package. Rows are the different tasks; values are the frequencies of the tasks done: by the wife only alternatively by the husband only or jointly.

	Wife	Alternating	Husband	Jointly
Laundry	156	14	2	4
Main_meal	124	20	5	4
Dinner	77	11	7	13
Breakfeast	82	36	15	7
Tidying	53	11	1	57
Dishes	32	24	4	53
Shopping	33	23	9	55
Official	12	46	23	15
Driving	10	51	75	3
Finances	13	13	21	66
Insurance	8	1	53	77
Repairs	0	3	160	2
Holidays	0	1	6	153

Use the data to answer the following questions:
  i.   Use the contingency table and inspect and interpret the row and column profiles (3 marks)
  ii. Evaluate whether there is a significant dependence between row and column categories   (2 marks)
  iii. When is correspondence analysis considered for multivariate data? (2 marks)
  iv. What is the significance of eigen values in correspondence analysis? (2 marks)
  v.   What is the proportion of variability accounted for by the first and second dimension?   (3 marks)
  vi. Sketch a scree plot for the correspondence analysis and interpret   (2 marks)
  vii. With the help of appropriate contribution biplots, give detailed discussion on how the various household tasks are dependence on the household member(s) (6 marks)

QUESTION THREE (20 Marks)

The data set KE-Arrests (https://drive.google.com/drive/folders/1CZAS5ZgaTLAZufBxg4xBT8I4zspIpiv0?usp=sharing) contains a hypothetical/simulated statistic, in arrests per 100000 residents for assault, murder, and rape in each of the 47 counties in the year 2018. Also given is the percent of the population living in urban areas. Use R to solve this question.

    a) Obtain the variables summary statistics (means and standard deviations)   (2 marks)

    b) What is the proportion of missing data in each of the variables (assault, murder, rape, and urbanpop)   (2 marks)

    c) Develop a correlation matrix involving the above variables. Provide, the matrix. Also, indicate which two variables have strongest correlation.   (3 marks)

    d) With *prcomp* function, develop a principal component analysis to the hypothetical KEarrest data. Give a rotation summary of the principal components. (4 marks)

    e) Give the contributions of the components in terms of the amount of variance explained. Discuss the result.   (3 marks)

    f) Sketch a scree plot for the PCA with x-axis denoting principal component and y-axis denoting proportion of variance explained.   (2 marks)

    g) Create 2 principal components and analyse performance of various counties with regards to the variables (Hint: use *biplot* function in R)   (4 marks)

QUESTION FOUR (20 Marks)

An ecologists measured x_1=the tail length (in mm) and x_2=wing length (in mm) for a sample of 45 female hook-billed kites (single multivariate sample). The data is given in the table below and also available as ecology data here (https://drive.google.com/drive/folders/1CZAS5ZgaTLAZufBxg4xBT8I4zspIpiv0?usp=sharing)

x_1	x_2	x_1	x_2	x_1	x_2
191	284	186	266	173	271
197	285	197	285	194	180
208	288	201	295	198	300
180	273	190	282	180	272
180	275	209	305	190	292
188	280	187	285	191	286
210	283	207	297	196	285

196	288	178	268	207	286
191	271	202	271	209	303
179	257	205	285	179	261
208	289	190	280	186	262
202	285	189	277	174	245
200	272	211	310	181	250
192	282	216	305	189	262
199	280	189	274	188	258

Required:

i.   Is the bivariate normal distribution a viable population model for this data? Explain   (2 marks)
ii.   Obtain variance-covariance matrix   (2 marks)
iii. Covert the var-cov matrix to a correlation matrix, calculate the eigen values and eigen vectors and verify that the eigen vectors are orthogonal   (4 marks)
iv.   Obtain the quadratic form of the covariance matrix   (4 marks)
v.   Suppose it is known that $\mu_1 = 190mm$ and $\mu_2 = 275mm$ for males birds, are these plausible values for μ_1 and μ_2 for female birds? Explain   (4 marks)
vi.   Find the 95% confidence ellipse for the population mean μ_1 and μ_2.   (4 marks)

QUESTION FIVE (20 Marks)

Use the data provided (also available here
https://drive.google.com/drive/folders/1CZAS5ZgaTLAZufBxg4xBT8I4zspIpiv0?usp=sharing) below relating to the production of plastic film reported in Krzanowski (2000). Tear, gloss and opacity are measures of manufactured films. The information on the rate of extrusion, and the amount of additive used are provided as experimental factors.

production of plastic films			experimental units	
tear	gloss	opacity	rate	additive
6.5	9.5	4.4	Low	Low
6.2	9.9	6.4	Low	Low
5.8	9.6	3.0	Low	Low

6.5	9.6	4.1	Low	Low
6.5	9.2	0.8	Low	Low
6.9	9.1	5.7	Low	High
7.2	10.0	2.0	Low	High
6.9	9.9	3.9	Low	High
6.1	9.5	1.9	Low	High
6.3	9.4	5.7	Low	High
6.7	9.1	2.8	High	Low
6.6	9.3	4.1	High	Low
7.2	8.3	3.8	High	Low
7.1	8.4	1.6	High	Low
6.8	8.5	3.4	High	Low
7.1	9.2	8.4	High	High
7.0	8.8	5.2	High	High
7.2	9.7	6.9	High	High
7.5	10.1	2.7	High	High
7.6	9.2	1.9	High	High

a) Conduct a multivariate analysis of variance and obtain the summary statistics (4 marks)

b) Obtain three matrices of interest in MANOVA

i. Total SSP (T) (4 marks)

ii. Between – group SSP (W) (3 marks)

iii. Within-group SSP (W) (3 marks)

iv. Wilk's Lambda $\frac{|W|}{T}$ (3 marks)

v. Examine the various group means (3 marks)