Kasarani Campus
Off Thika Road
Tel. 2042692 / 3
P. O. Box 49274, 00100
NAIROBI
Westlands Campus
Pamstech House
Woodvale Grove
Tel. 4442212
Fax: 4444175

KIRIRI WOMENS' UNIVERSITY OF SCIENCE AND TECHNOLOGY

UNIVERSITY EXAMINATION, 2022/2023 ACADEMIC YEAR

SECOND YEAR, SECOND SEMESTER EXAMINATION

FOR THE DEGREE OF BACHELOR OF SCIENCE

(MATHEMATICS AND COMPUTER SCIENCE)

Date: 2ND August, 2022
Time: 11.30am –1.30pm

KMA 208 - COMPUTER INTERACTIVE STATISTICS

INSTRUCTIONS TO CANDIDATES

ANSWER **QUESTION ONE (COMPULSORY)** AND **ANY OTHER TWO** QUESTIONS

QUESTION ONE (30 MARKS)

a) Explain what the following functions does
   i) cbind( ). (1 mark)
   ii) rbind( ). (1 mark)
   iii) seq (10, 20, 2). (1 mark)
   iv) rep(10, 5). (1 mark)
   v) table( ). (1 mark)

b) The data below relate to the score of 14 students in a CAT

   10, 25, 15, 17, 20, 9, 17, 15, 15, 16, 12, 28, 11,15

   Write a well commented R program that
   i) Input and print the score. (3 marks)
   ii) Compute the mean, median, 6th decile, 72nd percentile and variance for the data. (5 marks)
   iii) Computes harmonic mean. (2 marks)

c) Write an R program that imports data from excel file named **Score.xlsx** in a forder named **Exams** in drive **D.** (3 marks)

d) The number of cars passing through the university Gate is known to be Poisson distributed with parameter $\lambda = 10$ per hour. Write an R program that;
   i) Simmulate 200 of such data. (2 marks)
   ii) Computes the probability that 3 to 6 cars passes through the gate in one hour. (2 marks)
   iii) The cummulative distribution value at x=5. (2 marks)

e) Write an R program that determine the solution to the systems of linear equations given below, hence determine the output by method substitution.

$$3X + 2Y = 21$$

$$5X - 3Y = 16$$

(6 marks)

QUESTION TWO (20 MARKS)

To test on the effect of smoking and biking on heart disease, a sample of 498 individuals were tasted. The summary of a multiple linear regression model relating heart disease (Y) to biking (X_1) and smoking (X_2) are as in figure below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.984658   0.080137  186.99   <2e-16 ***
biking      -0.200133   0.001366 -146.53   <2e-16 ***
smoking      0.178334   0.003539   50.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.654 on 495 degrees of freedom
Multiple R-squared:  0.9796,    Adjusted R-squared:  0.9795
F-statistic: 1.19e+04 on 2 and 495 DF,  p-value: < 2.2e-16
```

a) Write down the model for the data. (3 marks)

b) Estimate the chance of a heart disease to a patient who bikes twice and doesn't smoke. (2 marks)

c) Write the R code that predicts the value obtained in b). (2 marks)

d) Using significant codes provided in the figure, test the significance of the model parameters at $\alpha = 0.05$ level of significance. (3 marks)

e) Test whether the overall model is significantly fitting the data or not at $\alpha = 0.05\%$ level of significance. (3 marks)

f) What percentage of the variation in the heart disease is explained by biking and smoking? (3 marks)

g) Assuming 8 arbitrary values of heart disease, biking and smoking, write R codes that produced the above figure. (4 marks)

QUESTION THREE (20 MARKS)

The following data shows the product price and its lifetime.

Lifetime (yrs.)	1	5	4	2	6	3
Price ($)	79	160	125	105	214	103

Write an R program that

a) Computes
   i) Pearson's correlation coefficient. (3 marks)
   ii) Spearman correlation coefficient. (2 marks)

b) Plots the scatter plot labeling both axes appropriately and give the title of the graph as **"Price ($) Vs Lifetime (yrs)".** (4 marks)

c) Add the line of the best fit to the scatter plot. (2 marks)

d) Determine the output of (a) (both (i) and (ii)) by calculations and comment on the results. (9 marks)

QUESTION FOUR (20 MARKS)

a) Let $X = (1, 4, 3, -1, 14, 17, -2, 5)$ and $Y = (2, 4, 7, 2, 5, 8, 12)$. Write R program that;

i) Finds the cumulative sum of values of X. (2 marks)

ii) Determines the cumulative product of values of Y. (2 marks)

iii) Assigns the 1st, 3rd, 6th and 7th values X to a new variable Z. (2 marks)

iv) Obtains the cumulative product of 2nd 3rd and 5th values of Y. (2 marks)

b) Determine the outputs of (a) (i) through (iv) by calculation. (4 marks)

c) Suppose that the score of the of students in a "KMA 208: Computer Interactive Statistics" is known to be normally distributed with mean 50 and standard deviation of 15. Write an R program that;

i) Simulate the score of 1000 students for the unit. (2 marks)

ii) Plot a histogram with 10 breaks labeling the axes and header as

Horizontal axis as "score"

Vertical axis as "Number of Students"

Header as "KMA 208: Computer Interactive Statistics" (4 marks)

iii) Add a normal density to the plot in (ii). (2 marks)

QUESTION FIVE (20 MARKS)

a) The following data show the scores of students in four different groups

A	60	30	80	45	30	55
B	77	40	63	50		
C	90	38	62	67	73	
D	82	48	50			

I) Write a well commented program in R that uses data to compute the following parameter estimate:

i) $$CS = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2 + (n_4-1)S_4^2}{n_1+n_2+n_3+n_4-4}}$$

ii) $$CM = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + n_3\overline{X}_3 + n_4\overline{X}_4}{n_1+n_2+n_3+n_4}$$

Where

$n_i =$ Sample size for i[th] branch, for $i = 1, 2, 3, 4$

$\overline{X}_i =$ Sample mean for i[th] branch, for $i = 1, 2, 3, 4$ (5 marks)

$S_i^2 =$ Sample variance for i[th] branch, for $i = 1, 2, 3, 4$

II) Hence determine the output of the estimates by calculation. (7 marks)

b) A matrix A is given by

$$A = \begin{pmatrix} 405.00 & -274.75 & 382.25 \\ -274.75 & 382.25 & -155.70 \\ 382.25 & -155.70 & 457.70 \end{pmatrix}$$

Write down an R program that

i) Enter and print matrix A. (3 marks)

ii) Determines the inverse and transpose of A. (3 marks)

iii) Obtains a diagonal matrix D of eigenvalues of A. (2 marks)