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

FOR THE DEGREE OF BACHELOR OF MATHEMATICS

(SPECIAL EXAMINATION)

KMA 208: COMPUTER INTERACTIVE STATISTICS

DATE: 3RD DECEMBER 2024
TIME: 2:30PM-4:30PM

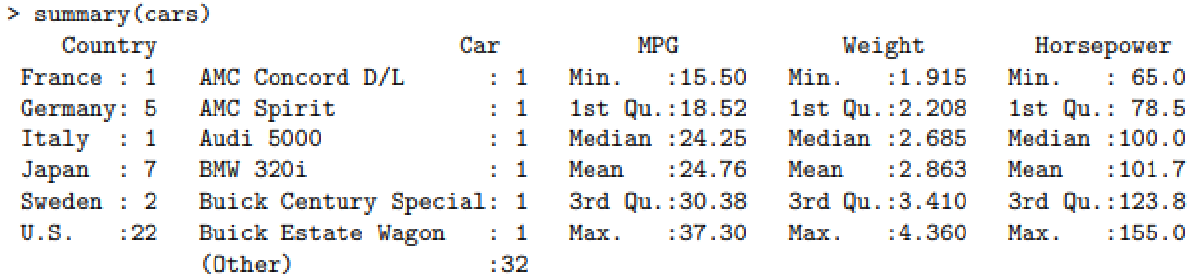
INSTRUCTIONS TO CANDIDATES

ANSWER QUESTION ONE (COMPULSORY) AND ANY OTHER TWO QUESTIONS

QUESTION ONE: COMPULSORY (30 MARKS)

(a) Discuss the data import process in R language. **(3 Marks)**

(b) Consider a data frame called cars:



```
> summary(cars)
   Country                       Car           MPG             Weight          Horsepower
 France : 1   AMC Concord D/L       : 1   Min.   :15.50   Min.   :1.915   Min.   : 65.0
 Germany: 5   AMC Spirit            : 1   1st Qu.:18.52   1st Qu.:2.208   1st Qu.: 78.5
 Italy  : 1   Audi 5000             : 1   Median :24.25   Median :2.685   Median :100.0
 Japan  : 7   BMW 320i              : 1   Mean   :24.76   Mean   :2.863   Mean   :101.7
 Sweden : 2   Buick Century Special: 1   3rd Qu.:30.38   3rd Qu.:3.410   3rd Qu.:123.8
 U.S.   :22   Buick Estate Wagon    : 1   Max.   :37.30   Max.   :4.360   Max.   :155.0
              (Other)               :32
```

(i) Write an R program to plot MPG on the y-axis and Horsepower on the x-axis, using a different color for each level of Country **(2 Marks)**

(ii) Write an R program that will show the row number of the observation with the with the highest ratio of MPG to weight. **(2 Marks)**

(c) Rose has kept a record of the number of times she had morning jog for the last 9 days. The data below shows the times in minutes.

20,17,16,22,24,21,15,17,22

(i) Write an R code for entering this data in R to a vector named "jog" **(1 Mark)**

(ii) Write an R code for getting mean, the longest jog time and the lowest jog time and give expected results **(3 Marks)**

(iii)She realizes that 24 was a mistake and should have been 18. Write an R code that will fix this. **(1 Mark)**

(iv)Write an R code which shows the number of times Rose jogged 19 minutes or more **(1 Mark)**

(d) Construct a matrix A with values 10, 20, 30, 50 in column 1, values 1, 4, 2, 3 in column 2 and values 15, 11, 19, 5 in column 3, i.e. a 4 × 3 matrix. Also construct a vector B with values 2.5, 3.5, 1.75. Check your results to ensure that they are correct. Combine A and B into a new matrix C using rbind(). **(5 Marks)**

(e) Simulate a sample of 100 random data points from a normal distribution with mean 100 and standard deviation 5, and store the result in a vector. Plot a histogram and a boxplot of the vector you just created. **(5 Marks)**

(f) Write functions tmpFn1 and tmpFn2 such that if xVec is the vector $(x_1, x_2, \ldots, x_n)$, then tmpFn1(xVec) returns the vector $(x_1, {x_2}^2, \ldots, {x_n}^n)$ and tmpFn2(xVec) returns the vector $(x_1, \frac{{x_2}^2}{2}, \ldots, \frac{{x_n}^n}{n})$. **(4 Marks)**

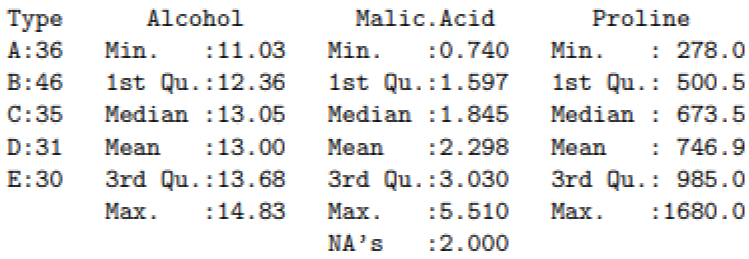
(g) Create the following matrix B with 15 rows

$$B = \begin{pmatrix} 10 & -10 & 10 \\ 10 & -10 & 10 \\ \cdots & \cdots & \cdots \\ 10 & -10 & 10 \end{pmatrix}$$

Calculate the 3 × 3 matrix $B^T B$ **(3 Marks)**

QUESTION TWO: (20 MARKS)

(a) Consider a data frame called wine, which contains information about the chemical composition of different types of wines. Here is some information about the data frame



```
Type      Alcohol         Malic.Acid        Proline
A:36   Min.   :11.03   Min.   :0.740   Min.   : 278.0
B:46   1st Qu.:12.36   1st Qu.:1.597   1st Qu.: 500.5
C:35   Median :13.05   Median :1.845   Median : 673.5
D:31   Mean   :13.00   Mean   :2.298   Mean   : 746.9
E:30   3rd Qu.:13.68   3rd Qu.:3.030   3rd Qu.: 985.0
       Max.   :14.83   Max.   :5.510   Max.   :1680.0
                       NA's   :2.000
```

(i) Write an R program that willl calculate the median of Alcohol and Malic.Acid for each Type of wine. **(2 Marks)**

(ii) Write an R program to count the number of observations with Alcohol greater than 13 and Proline less than 650. **(2 Marks)**

(iii)If you were reading this data from a comma-separated file, what option would be passed to read.csv to ensure that Type was read as a character variable, not a factor? **(2 Marks)**

(iv)Write an R program to produce a barplot showing the number of wines of each type in the data frame. **(2 Marks)**

(b) We type the following in R:

```
> theta <- c(8, 6, 4, 2)
 > rho <- c(0, 1)
> delta <- c(TRUE,TRUE,FALSE,TRUE,FALSE)
 > phi <- seq(from=0, to=8, length=5)
```

Given the assignments above, what is the output of the following commands?

(i) theta [1: 3] **(1 Mark)**

(ii) theta [-2] **(1 Mark)**

(iii) theta-rho **(2 Marks)**

(iv) 3-theta/seq(from=4, to=l) **(2 Marks)**

(c) Explain what each line of the R code does and give the expected outputs for each

(i) K<-cbind(L=1:3, M=4:6, N=3) **(3 Marks)**

(ii) B<-rbind(c(1,2,3),5:3,c(100,20,70),(11:13)) **(3 Marks)**

QUESTION THREE: (20 MARKS)

(a) The following data represents alcohol concentration in the blood sample of 10 drivers along a certain road as well as their driving speeds

Acohol Conc.	1.55	1.71	1.39	1.15	1.33	1.00	1.68	1.76
Speed(Km/h)	61	60	100	93	78	80	99	120

Required:

Analyze the above data using regression. Write the basic syntax for the regression analysis in R. Write a well commented program in R that does the following

(i) Reads in data **(3 Marks)**

(ii) Fits a linear model to the data but provides no further statistical information to the model **(2 Marks)**

(iii)Provides a complete statistical summary of the model **(2 Marks)**

(iv)Check whether the observed data meets our model assumptions **(3 Marks)**

(v) Visualize the results of your simple linear regression. **(2 Marks)**

(vi)Add the linear regression line to the plotted data. **(3 Marks)**

(b) Write a custom function which will replace all the missing values in the vector data<-c(12,25,NA,89,78,NA,36,14,26,NA) with the mean of values. **(5 Marks)**

QUESTION FOUR: (20 MARKS)

(a) Given the following two matrices

$$A = \begin{pmatrix} 0 & 4 & -6 \\ 5 & 6 & 9 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & 4 & 7 \\ 5 & 5 & 8 \\ 5 & 2 & 2 \end{pmatrix}$$

Write the R program that does the following

(i) Reads and display the two matrices A and B **(2 Marks)**

(ii) Adds the two matrices **(2 Marks)**

(iii)Transpose of A× B **(4 Marks)**

(b) Consider the following system of linear equation, solve for x1 and x2 using R **(4 Marks)**

3x1 + 4x2 =4

x1+x2 =2

(c) Consider the following vector:

```
> text = c('cat 122','dog 213','721 chicken','fish 42','893 duck')
```

Use regular expressions to answer the following questions:

(i) Write an R program to create a vector like text, with the number in each element appearing before the animal name **(2 Marks)**

(ii) Write an R program to create a vector containing just the animal names in text. **(2 Marks)**

(iii)Write an R program to produce a vector containing the position of the blank in each element of text. **(2 Marks)**

(iv)Write an R program to remove the first three characters in each of the elements of text **(2 Marks)**

<u>QUESTION FIVE: (20 MARKS)</u>

(a) Calculate the following $\sum_{i=1}^{25}(\frac{2^i}{i} + \frac{3^i}{i^2})$ **(4 Marks)**

(b) Consider a data frame called trees

```
> summary(trees)
    Girth          Height       Volume
 Min.   : 8.30   Min.   :63   Min.   :10.20
 1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
 Median :12.90   Median :76   Median :24.20
 Mean   :13.25   Mean   :76   Mean   :30.17
 3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
 Max.   :20.60   Max.   :87   Max.   :77.00
```

(i) Write a summary statistic of the variables Girth, Height and Volume. **(4 Marks)**

(ii) Visualize the distribution of Girth with a stem-and-leaf

```
The decimal point is at the |
  8 | 368
 10 | 57800123447
 12 | 099378
 14 | 025
 16 | 03359
 18 | 00
 20   6
```

Does the distribution appear symmetric? **(2 Marks)**

(c) Consider the data

```
workshop <- c("R", " SPSS ", NA , " SPSS ", " STATA ", " SPSS ")
gender <- factor (c(" Female ", " Male ", NA , " Female ", " Female ",
" Female ") )
    q1 <- c(4 , 3 , 3 , 5 , 4 , 5)
    q2 <- c(3 , 4 , 2 , 4 , 4 , 4)
     q3 <- c(4 , 3 , NA , 5 , 3, 3)
    q4 <- c(5 , 4 , 3 , 3 , 4 , 5)
    df <- data . frame ( workshop , gender , q1 , q2 , q3 , q4 )
```

(i) Create a dataframe consisting of only the first two columns. **(1 Mark)**

(ii) Create a dataframe consisting of only the first and last row. **(1 Mark)**

(iii)Create a dataframe called df2 where every entry in the q3 and q4 columns is 0. **(2 Marks)**

(iv)Sort df by gender. **(1 Marks)**

(v) Does df have any duplicate rows? **(1 Marks)**

(d) Write a function to generate n random numbers from the distribution with density

$f(x) = 3x^2, 0 < x < 1$ **(4 Marks)**