MERU UNIVERSITY OF SCIENCE AND TECHNOLOGY

P.O. Box 972-60200 – Meru-Kenya.
Tel: +254(0) 799 529 958, +254(0) 799 529 959, +254 (0)712 524 293
Website: www.must.ac.ke Email: info@mucst.ac.ke

UNIVERSITY EXAMINATIONS 2024/2025

THIRD YEAR, FIRST SEMESTER EXAMINATION FOR THE DEGREE OF BACHELOR OF DATA SCIENCE

CDS 3300: STATISTICAL INFERENCE AND MODELLING IN R

DATE: JANUARY 2025 **TIME: 2 HOURS**

INSTRUCTIONS: Answer Question ONE and any other TWO questions.

QUESTION ONE (30 MARKS)

a) Explain what an Integrated Development environment (DE) (2 Marks)

b) Describe the typical layout of the RStudio interface (8 Marks)

c) Define a mathematical model and provide two examples of where mathematical models might be used in a business context (6 Marks)

d) Write a simple R script that generates a sequence of numbers from 1 to 10 and prints each number to the console (2 Marks)

e) Write a function called 'calculate area' that takes the radius of a circle as an argument and returns its area. Use pi from the R environment in your calculation. (3 Marks)

f) Explain the concept of resampling. What are the primary differences between bootstrapping and cross-validation? Provide examples of when to use each technique (9 Marks)

QUESTION TWO (20 MARKS)

a) Consider a customer service desk that receives customers at a rate of 10 per hour (assume arrivals follow a Poisson distribution). Write an R code snippet to simulate the number of customer arrivals in an 8-hour day. Based on your simulation, write an R code to calculate the mean and variance of the customer arrivals in one hour. (6 Marks)

b) Write an R script to simulate rolling a fair six-sided die 1000 times (4 Marks)

c) Explain how simulations can help to model real-world processes. Give an example of a real-world process that could benefit from a simulation (10 Marks)

QUESTION THREE (20 MARKS)

Explain three assumptions in regression modelling (6 Marks)

Give five advantages of regression modelling (5 Marks)

The following regression output relates to an investigation on relationship between two variables. y: the dependent variable (e.g., house prices), x: the independent variable (square footage)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  124.5702    49.2510   2.529   0.0132 *
x              0.4934     0.0225  21.922   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 97.15 on 98 degrees of freedom
Multiple R-squared:  0.8304,    Adjusted R-squared:  0.8285
F-statistic: 480.6 on 1 and 98 DF,  p-value: < 2.2e-16
```

Provide a detailed interpretation of the output (9 Marks)

QUESTION FOUR (20 MARKS)

a) You're given a dataset containing customer purchase data. Describe how you would perform an EDA to understand (give examples of RStudio codes)

   i. Customer buying patterns (4 Marks)
   ii. Purchase anomalies (e.g., unusually high spending by certain customers) (4 Marks)
   iii. Data distributions for purchase amounts (4 Marks)

b) What are the benefits of using functions in programming? (8 Marks)

QUESTION FIVE (20 MARKS)

a) What is a data frame in R? Explain how it is different from a matrix (5 Marks)
b) Write an R code to create a data frame named students with the following columns and data:

   Name: "John", "Lisa", "Tom"

   Age: 20, 22, 23

   Grade: "A", "B", "A" (2 Marks)

c) Define a vector in the context of programming. Why are vectors useful in data manipulation and analysis? (8 Marks)
d) if you have two vectors, a <- c(3, 5, 7) and b <- c(2, 4, 6), what will the result of a + b and a * b be? Explain how R handles element-wise operations. (5 Marks)