JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF BIOLOGICAL, PHYSICAL, MATHEMATICS AND ACTUARIAL SCIENCES
UNIVERSITY ASSESSMENT FOR CBET DIPLOMA IN APPLIED STATISTICS

2nd Year 1st SEMESTER 2023/2024 ACADEMIC YEAR
MAIN REGULAR

COURSE CODE: WAB 2214

COURSE TITLE: STATISTICAL COMPUTING I

EXAM VENUE: **STREAM: (Dip. Applied Statistics)**

DATE: **EXAM SESSION: Sep-Dec 2023**

TIME: 3.00 HOURS

Instructions:

i. Answer questions one and any other three.
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. Candidates are advised to carry their personal computers with **R** and **ISwR** package installed beforehand.

SECTION A (40 marks)

QUESTION ONE (40 MARKS)

a) In R, how are missing values are represented? (1 marks)
b) How can one save and or write data in csv and stata file formats with R? (1 marks)
c) If x is a factor with n levels and y is a length n vector, what happensif you compute y[x]? (1 marks)
d) Write the logical expression to use to extract girls between 7 and 14 years of age in the *juul* data set. (2 marks)
e) List four advantages of R programming software (4 marks)
f) If you make a plot like plot(rnorm(10),type="o") with overplotted lines and points, the lines will be visible inside the plottingsymbols. How can this be avoided? (1 marks)
g) Provide R-codes used to generate random numbers from the following distributions
   i. Normal (1 mark)
   ii. Exponential (1 mark)
   iii. Gamma (1 mark)
   iv. Poisson (1 mark)
   v. Binomial (1 mark)
h) In the data set *vitcap*, use a *t* test to compare the vital capacity for the two groups. Calculate a 99% confidence interval for the difference. The result of this comparison may be misleading. Why? (2 marks)

i) Explain the following giving examples in each case

i. Nominal and ordinal scales of measurements (2 marks)

ii. Numeric and character data types (2 marks)

j) Perform the analyses of the react and *vitcap* data using nonparametric techniques. (2 marks)

k) Perform graphical checks of the assumptions for a paired *t* test in the *intake* data set. (1 marks)

QUESTION TWO (20 Marks)

The following data represent the body-breadth (x, in cm) and body-weight (y, in cm) of 14 randomly selected sea fishes.

x	0.5	0.6	0.8	0.4	0.5	0.7	1.0	1.0	0.6	0.7	1.5	0.5	0.5	0.6
y	10	15	25	12	15	14	25	28	18	20	40	18	15	20

i. Draw a scatter diagram providing an R-code (3 marks)

ii. At α=0.05, examine whether body-weight and body-breadth of fishes are significantly correlated (5 marks)

iii. Provide an R-code to obtain spearman and Pearson correlation coefficients (2 marks)

iv. Fit a simple linear regression of y on x extracting regression slope and intercept (8 marks)

v. Write an R-code that gives a summary and ANOVA of the fit in (iv) (2 marks)

QUESTION THREE (20 Marks)

(a) The following data are the random sample of observations recorded from an industry producing juice per hour in different days

x_i: 50,55,62,67,45,68,70,62,73,64,75,55,50,68,64,60,66,60,56,59,60,60,63,67,66,68,70,65,54,55,70,66,67

i. Plot a boxplot to explore the above data (5 marks)

ii. Under the assumption xi $\sim N(\mu, \sigma^2)$, testthe significance of H0: $\mu = 60$ against H1: $\mu \neq \mu_0$at $\alpha = 0.05$ (5 marks)

iii. Provide R-codes (i) and (ii) (2 marks)

(b) If the number of boys and girls who are regular in their exercises is distributed as given below, it is required to test whether there is statistical difference in the exercise habits between boys and girls.

Habit	Boys	Girls	Total
Exercise regularly	2	8	10
Do not exercise regularly	10	4	14
Total	12	12	24

i) Use an appropriate Chi-square test to assess whether exercise regularly is associated with sex at 95% confidence level. (5 marks)

ii) Provide an R code that can generate the results in (i). (3 marks)

QUESTION FOUR (20 MARKS)

a) Calculate the probability for each of the following events:

i. A standard normally distributed variable is larger than 3. (3 marks)

ii. A normally distributed variable with mean 35 and standard deviation 6 is larger than 42. (3 marks)

iii. Getting 10 out of 10 successes in a binomial distribution with probability 0.8. (3 marks)

iv. *X* <0.9 when *X* has the standard uniform distribution. (3 marks)

v. *X* >6.5 in a χ2 distribution with 2 degrees of freedom. (3 marks)

b) For a disease known to have a postoperative complication frequency of 20%, a surgeon suggests a new procedure. He tests it on 10 patients and there are no complications. What is the probability of operating on 10 patients successfully with the traditional method? (5 marks)

QUESTION FIVE (20 MARKS)

Suppose that to simulate from the following linear model

$y = \beta_0 + \beta_1 x + \varepsilon$ where $\varepsilon \sim N(0, 2^2)$. Assume $x \sim N(0, 1^2)$, $\beta_0 = 0.5$ and $\beta_1 = 2$

set.seed(20); Let: x=rnorm(100); ε =rnorm(100,0,2); $y = 0.5 + 2 * x + \varepsilon$

Given that x and ε are normally distributed

i. Obtain the following for both x and y variables and provide an interpretation for each case
   a) mean (2 marks)
   b) variance (2 marks)
   c) minimum (1 mark)
   d) maximum (1 mark)
   e) range (1 mark)

ii. Plot a scatter diagram between x and y and interpret accordingly. (2 marks)

iii. Compute the correlation coefficient
   a) Pearson (2 marks)
   b) Spearman (2 marks)

   Interpret them accordingly.
   c) Compare and contrast the correlation coefficients in a) and b) (2 marks)
   d) Using Pearson procedure, test the hypothesis that correlation coefficient is significant at 95% confidence level (3 marks)
   e) Interpret the slope parameters tests their significance (2 marks)

QUESTION SIX(20 MARKS)

i. From Agresti(2007) p39

m=as.table(rbind(c(762, 327, 468), c(484, 239, 477)))

dimnames(m) <- list(gender=c("F", "M"), party=c("Democrat","Independent","Republican"))

   a) Obtain a summary of the Chi-square test (3 marks)
   b) Provide a procedure(R-code) of obtaining observed counts (same as m) (3 marks)
   c) Expected counts under the null (3 marks)
   d) Pearson residuals (3 marks)
   e) Standardized residuals (2 marks)
   f) Chi-square statistic (1 mark)
   g) Chi-square p-value (1 mark)
   h) Test the hypothesis whether gender and party are independent (2 marks)
   i) Interpret the findings in h) (2 marks)