JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF BIOLOGICAL, PYSICAL, MATHEMATICS AND ACTUARIAL SCIENCES
UNIVERSITY EXAMINATION FOR DEGREE OF BACHELOR OF SCIENCE IN ACTUARIAL SCIENCE

2nd Year 1st SEMESTER 2021/2022 ACADEMIC YEAR
MAIN REGULAR

COURSE CODE: WAB 2209

COURSE TITLE: STATISTICAL COMPUTING I

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. Computations and data analysis to be done with R statistical software.

QUESTION ONE

i. Use the hsb data provided to answer the following questions.

```
hsb2 <- read.table("https://stats.idre.ucla.edu/wp-
content/uploads/2016/02/hsb2-1.csv", header=T, sep=",")
```

a. Provide the mean and standard deviation of socst by prog, explain the result (1 mark)
b. Provide the mean and standard deviation of math by race (1 mark)
c. Develop boxplot of performance in math by ses and interpret the result (1 mark)

ii. Develop a matrix using elements provided below:

```
set.seed(1234)
x <- matrix(rnorm(30, 1), ncol = 5)
y <- c(1, seq(5))
```

a. Give a resulting matrix when x and y are combined into one matrix X. Round off to the nearest 2 decimal places. (1 mark)
b. convert x into a data frame called x.df. Provide x.df and round off to the nearest 2 decimal places. (1 mark)
c. X transpose (1 mark)
d. Diagonal matrix of X (1 mark)
e. Determinant and inverse of X (1 mark)
f. Cholesky factorization of X which returns the upper triangular factor such that $X^1X=A$ (1 mark)

iii. Can you write and explain some of the most common syntax in R (1 marks)
iv. How would one list pre-loaded datasets in R? (1 marks)
v. List any four help packages most recommended for R users (2 marks)
vi. What are the different data types/objects in R (2 marks)
vii. Explain use of "source" function in R (2 marks)
viii. Write an R function that returns mean, standard deviation, minimum, maximum, and range of a vector. (2 marks)
ix. How would one install a package in R? (1 marks)

QUESTION TWO

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 THREE

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 (2 marks)
   b) Provide a procedure(R-code) of obtaining observed counts (same as m) (2 marks)
   c) Expected counts under the null (2 marks)
   d) Pearson residuals (2 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 (1 marks)
i) Interpret the findings in h) (2 marks)

ii. Re-write the data as a dataframe with three columns (Freq(counts), gender and party
   a) Evaluate whether gender and party have any relationship using loglinear model (2 marks)
   b) Compare the results in i.h) and ii.a) (1 marks)
   c) When is a Chi-square and Fisher's exact test used? (1 marks)

QUESTION FOUR

Use high school and beyond (hsb) dataset. The data contains 200 observations from a sample of high school students with demographic information about the students such as gender(female), socio-economic status(ses) and ethnic background(race). It also contains a number of scores on standardized tests including tests of reading(read), writing(write), mathematics(math), science and social studies(socst)

i. Use an appropriate test to assess whether the average performance in math=55. Provide a summary output and interpret the result. (3 marks)

ii. Test whether the proportion of males and females are equal. (3 marks)

iii. Generate a variable that sums all the performance scores (read, write, math, science and socst). Call it totalscore
   a) Develop a histogram with a density curve overlaid. Interpret. (3 marks)
   b) Assess the normality of the totalscore with a stem and leaf plot (3 marks)
   c) Evaluate the distribution of totalscore across various races with a boxplot. Interpret the boxplots accordingly (3 marks)
   d) When is a scatter diagram used? (2 marks)
   e) Use an appropriate graph to assess the direction of the relationship between totalscore and math. (3 marks)

QUESTION FIVE

During a recent agricultural show, two judges were assigned the duty of awarding marks for a given category of produce exhibited at different stands. Each judge was under strict instructions to give their own honest opinion on the exhibition. The marks awarded by each of the two judges to 8 exhibitors were recorded as follows

Judge 1	22	27	24	17	20	22	16	13
Judge 2	28	23	25	14	26	17	20	15

i. Calculate the Spearman's rank correlation between the marks awarded by the judges. (2 marks)
ii. Stating your hypotheses and using a 5% level of significance, interpret your results. (3 marks)
iii. Using appropriate test compare marks awarded by the two judges stating whether the marks as awarded were statistically different (4 marks)

Create a two-column data frame. Let the first column be name "judge" and second column be "marks"

i. Give the two types of data expressed in each of the two variables (2mark)
ii. Provide mean and standard deviation of marks awarded by judge. (2mark)
iii. Develop a scatter diagram overlay the best fitting line with R. (3mark)
iv. State the shape of the relationship displayed. (2mark)
v. Add the best fitting line to the scatter diagram and interpret. (2mark)
vi. Fit a regression equation of marks on the categorical variable (judge)
    - Write down the regression model. (2mark)
    - Interpret the regression parameter estimates. (2mark)
    - State the fit statistic and interpret. (1mark)
    - What is the relationship between R^2 and the Spearman's correlation coefficient obtained in 1(a)? (1mark)