TECHNICAL UNIVERSITY OF MOMBASA

FACULTY OF APPLIED AND HEALTH SCIENCES

DEPARTMENT OF MATHEMATICS & PHYSICS

UNIVERSITY EXAMINATION FOR:

THE DEGREE OF BACHELOR OF SCIENCE IN STATISTICS & COMPUTER SCIENCE

AMA 4320: **STATISTICAL MODELLING**

END OF SEMESTER EXAMINATION

SERIES:APRIL2016

TIME:2HOURS

DATE: Pick DateMay2016

Instructions to Candidates

You should have the following for this examination

-Answer Booklet, examination pass and student ID

This paper consists of **FIVE** questions. Attemptquestion ONE (Compulsory) and any other TWO questions.

Do not write on the question paper.

Question ONE

a) i) Write a R code that generates a vector with a sequence of numbers from 10 to 20 with 0.5 steps. Name your vector x. (3 marks)

ii) Using vector x above, what results would you get after running this code; (2 marks)
x[x>=15]

b) A lecturer would like to write R loop script that can grade students' performance. He has written the following loop script;

```
x <-60
if (x >=80){
    print('Excellent')
} else if (x<80 & x>=50){
    print('Pass')
} else {
    print('Fail')
}
```

i) Identify the conditional expression in the loop statement. (3 marks)

ii) List the results of the loop statement above. (1 marks)

iii) Edit the loop script to display grade for a student who scored 32 marks. (1 mark)

c) You are working on a directory with the following objects;

```
> dir()
[1] "autolab.dta"         "babies.csv"          "baby.sav"
[4] "binary.sav"          "faminc.dta"          "mydata.RData"
[7] "pair_data.sav"       "Pneumonia_data.sav" "regress.sav"
```

i) Write a code to read "faminc.dta" dataset into R. (3 marks)

ii) The code you used above depends on a certain R **Package**. Write a script to load the R package. (2 marks)

d) Given the below matrix, write R script to have the matrix in R; (5 marks)

$$\mathbf{C} = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix}$$

e) i) A researcher proposes that left handness is associated with risk of car accident. To prove his theory, he got data on whether a driver was left or right handed and if they had an accident or not. What statistical test would he use to test the null hypothesis of no association between right or left handness and risk of committing an accident? (2 marks)

iii) Suppose the dependent variable (accident) was labelled 'y' and the independent variable (handness) was labelled 'x'. Write a code in R to test the null hypothesis of no association between right or left handness and risk of committing an accident using the statistical test you named above. (3 marks)

f) **i)** Suppose you are interested in generating 1000 random numbers using the uniform distribution. Write a code in R that can generate the 1000 random numbers. Call the generated random numbers from uniform distribution y. (3 marks)

ii) Write a code to plot a histogram of vector y created above. (2 marks)

Question TWO

i) You are needed to solve the linear equations below in R.

3x+4y+3z=4
2x+y+2z=8
5x+4y+5z=7
Write a code to create the 3X3 matrix that includes the x, y and z data. Call the matrix 'y' (9 marks)

ii) Write a code to create the vector containing the results of the three equations (4, 8 and 7) call the vector x. (3 marks)

iii) Write a code that can generate an inverse of matrix 'y'. (3 marks)

iv) You run a code that produce the following results;

```
3 x 1 Matrix of class "dgeMatrix"
          [,1]
[1,]  3.363636
[2,] -4.318182
[3,]  1.863636
```

Write the code used to produce these results. (5 marks)

Question THREE

a) An author is interested in having many graphs on one screen to include in his manuscript for publication. You have been approached to help him out. You start out by writing the code below in R.

```
>split.screen(figs=c(2, 2))
```

Explain the use of the above R code. (5 marks)

The author had already written the below R codes to create the graphs.

```
> hist(weight)
> plot(weight~agemonth)
> hist(height)
> plot(height~agemonth)
```

b) Write a script to plot the R code that would generate all the above graphs in one screen. Add a title and label for the x-axis in every graph. Give the boxes in each plot different color. (15 marks)

Question FOUR

Grade point average (GPA) is a grading system used in some countries for admission to college. A Social scientist is interested in studying the effect of GPA on college admission. GPA is a continuous variable assumed to have a normal distribution ranging from 1 to 6. College admission is a binary variable coded as 1 for those admitted and 0 for those not admitted to college. The social scientist hired a statistician to help with the analysis. Below are the first lines codes in R written by the statistician to export data into R;

>data_binary<-read.spss("binary.sav", to.data.frame=TRUE)
>attach(data_binary)
>names(data_binary)

a) Explain the purpose of three lines of script above. (3 marks)
b) Name the appropriate statistical test that the statistician used for this analysis. (2 marks)
c) Given that, the binary admission variable was labelled 'admit' and the GPA grade variable was labelled 'gpa', write the code required to run the statistical test you listed above. (2 marks)
d) Below are the analysis results produced by the hired statistician;

```
data:  gpa by admit
t = -3.6379, df = 250.049, p-value = 0.0003339
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.22429214 -0.06673381
sample estimates:
mean in group 0 mean in group 1
       3.343700        3.489213
```

What was the GPA mean difference between the two groups (those admitted and those not admitted) and the 95% confidence interval of the difference? (6 marks)

e) Was there any evidence against the null hypothesis of no mean GPA difference between the two groups? (7 marks)

Question FIVE

Analysis of variance (ANOVA) is used to test group differences on the mean of a continuous variable divided up by a categorical variable with more than two levels.

a) State three assumptions underlying ANOVA. (6 marks)

A researcher believes that the level of haemoglobin among first, second and third year diploma students does not differ across the year of academic level. To test his hypothesis he sampled 912 students from first year to third year and obtained their haemoglobin levels. He then used the below R codes to run ANOVA;

>myModel<- aov(hg~class)
>anova(myModel)

b) From the codes above, identify the dependent and independent variables. (4 marks)

c) Here are the output of the analysis;

```
Analysis of Variance Table

Response: hg
           Df Sum Sq Mean Sq F value  Pr(>F)
class       1   27.5 27.5086         0.01798 *
Residuals 888 4347.4  4.8957
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F value has been omitted from the output. Compute the omitted F value. (3 marks)

d) Was the researcher right to believe that levels of haemoglobin didn't differ across the three years of study? What can you conclude from the results? (7 marks)