MACHAKOS UNIVERSITY

University Examinations 2019/2020 Academic Year

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND STATISTICS

THIRD YEAR SPECIAL /SUPPLEMENTARY EXAMINATION FOR

BACHELOR OF SCIENCE IN MATHEMATICS AND COMPUTER SCIENCE

BACHELOR OF SCIENCE IN STATISTICS AND PROGRAMMING

SST 301 PROGRAMMING LANGUAGE FOR STATISTICS I

DATE: 22/01/2021 **TIME: 8.30-10.30 AM**

INSTRUCTIONS

Answer question one and any other two Questions

QUESTION ONE (30 MARKS)

a) Explain briefly the following concepts as used in statistics

i. Hypothesis testing

ii. Homoscedasticity (2 marks)

b) Given the following vectors, vec1 $< -c(1,2,3,4,5)$, vec2 $< -c(1,3,45,67,89,32)$, write R output for the following functions

i. union(vec1,vec2)

ii. setdiff(vec1,vec2)

iii. setequal(vec1,vec2) (3 marks)

c) Write R output for the following

i. matrix(c(2,4,3,1,-1,6),nrow=2,ncol=3,byrow=TRUE) (2 marks)

ii. matrix(2:7,nrow=2) (2 marks)

iii. rep(c(1,3,5),c(3,5,7)) (2 marks)

iv. factor(c(1,0,1,0,0,0),levels=c(0,1),labels=c("girl","boy")) (2 marks)

d) Let $A = \begin{pmatrix} 10 \\ 20 \\ 5 \end{pmatrix}$ and $= \begin{pmatrix} -3 \\ 15 \\ -6 \end{pmatrix}$. Write R code to evaluate $A^2 + 2A + {}^{B}/_{9} + 3$ (3 marks)

e) Write down the simple linear regression model explaining all its components and the assumptions taken into account when defining the model. (6 marks)

f) Consider the data below on weekend sales of ice cream in Nairobi.

Temperature (X)	25	16	28	20	22	23	16	18
Sales(in 100 shillings) (Y)	125	79	140	103	111	115	80	91

Assume that the data is in a notepad file named *icedata* stored in directory E and folder myDocs in your computer. Write R code that does the following:

i. Reads in the data into R. (2 marks)

ii. Does a scatter plot of the data with correct labels for the X and Y axes and has the title 'SCATTER PLOT OF ICE CREAM SALES'. (3 marks)

iii. Fits a simple linear regression model. (2 marks)

iv. Superimposes the fitted regression line on the scatter plot. (1 mark)

QUESTION TWO (20 MARKS)

a) Given the table below,

Name	**Age**	**Gender**	**Points**
Alice	37	F	278
Paul	34	M	200
Jerry	26	M	157
Thomas	72	M	800
Mary	18	F	300
Linda	24	F	96

i. Create a data frame called shopping_points with the data in R. (5 marks)

ii. Subset the data for males into a data frame called mpoints. (1 mark)

iii. Extract the data for shoppers with more than 200 points and above the age of 30. (2 marks)

iv. Display the column with **Points** only (1 mark)

v. Display row four of the data frame (1 mark)

b) The data sales<-c(33,44,29,16,25,45,33,19,54,22,21,49,11,24,56) contain sales of milk in litres for 5 days in three different shops. The first three values are for shops 1,2 and 3 on Monday, the next three values are for the three shops on Tuesday and so on. Produce a statistical summary of the sales for each day of the week and for each shop using R. (5 marks)

c) Write a function in R with two input variables which are numerical. The function should add, subtract, multiply and divide the two variables and print as output all these. Further, an error message should be displayed if the input variable is not numerical or is zero. (5 marks)

QUESTION THREE (20 MARKS)

a) Briefly explain the difference between descriptive and inferential statistics. (2 marks)

b) A survey was done to examine the relationship between different continents and new cases of breast cancer. The null hypothesis stated that the means for the seven continents were equal while the alternative hypothesis stated that not all seven continents means are equal. ANOVA was used to test this hypothesis and the R output is as below.

```
                  Df Sum Sq Mean Sq F value Pr(>F)
mydata$continent   6  52531    8755   40.28 <2e-16 ***
Residuals        166  36083     217
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

i. Write the R command that produced the output above. (2 marks)

ii. Interpret the output. (3 marks)

iii. Explain what can be done to determine which groups are different from each other. Write R code that can be used to test this. (2 marks)

c) A multiple regression was carried out to find out whether the variables age and number of siblings (no_siblings) had an effect on the height of a child.

i. Write the R code for this. (1 mark)

ii. The following output was obtained

```
Residuals:
    Min      1Q  Median      3Q     Max
-47.808  -4.521   1.002  10.025  21.186

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   56.597     39.166   1.445   0.1823
age            1.303      1.712   0.761   0.4660
no_siblings   -6.415      3.464  -1.852   0.0971 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.59 on 9 degrees of freedom
Multiple R-squared:  0.2791,    Adjusted R-squared:  0.1189
F-statistic: 1.742 on 2 and 9 DF,  p-value: 0.2293
```

Write down the regression line. (1 mark)

iii. Write down the hypothesis that you would test to check whether the explanatory variables are contributing significantly in predicting the response variable. (2 marks)

iv. From the output above, what would be the result of your tests in iii above at 5% significant level? (3 marks)

v. What is the difference between the Multiple R squared and Adjusted R squared as given in R output? (2 marks)

vi. According to the output, would you say that the model is a good fit? Explain your answer. (2 marks)

QUESTION FOUR (20 MARKS)

a) Differentiate between the terms primary data and secondary data as used in statistics giving examples in each case. (4 marks)

b) Six people are asked to rate the performance of a product on a scale of 1-5, with 1 representing very poor performance and 5 representing very good performance. The following data on their responses 1,3,4,2,2,1 was obtained. Write R code to represent this data and give the expected output.

(5 marks)

c) Given the matrices $A = \begin{bmatrix} 2 & 4 \\ 5 & 8 \end{bmatrix}$ and $B = \begin{bmatrix} 3 & 2 \\ 6 & 7 \end{bmatrix}$, write R code to

i. Read in and display the matrices in R. (2 marks)

ii. Obtain the sum of the matrices and display the result in R. (2 marks)

iii. Obtain and display the transpose of the result in ii above. (1 mark)

d) A study was conducted to investigate the relationship between the price of a plot of land and the distance from the nearest tarmac road. A random sample of 12 plots was taken and the data relating the distance from the nearest tarmac road in hundred metres and the price of the plot in million Kenya shillings is as shown in the table below.

Plot	1	2	3	4	5	6	7	8	9	10	11	12
Distance	48	78	35	72	10	60	40	22	84	25	12	48
Price	1.2	0.4	1.4	0.8	1.8	0.6	1.5	1.2	0.8	1.5	1.6	1.2

i. Assume the data is stored as file *plot.txt* in directory E and folder RDocs in your computer. Read in the file into R with the headings. (2 marks)

ii. Write R code to calculate the mean and standard deviation of the two variables. (2 marks)

iii. Write R code to compute the Pearson's product moment co-efficient of correlation between the distance from the tarmac and the price of the plot. (2 marks)

QUESTION FIVE (20 MARKS)

a) Outline the procedure used for hypothesis testing. (4 marks)

b) Differentiate between Type I and Type II errors as used in hypothesis testing. (2 marks)

c) In order to increase the efficiency with which employees in a certain organisation can carry out a task, 5 employees are sent on a training course. The time in seconds to carry out the task both before and after the training course is given below for the 5 employees:

Employee	1	2	3	4	5
Before	42	51	37	43	45
After	38	37	32	40	48

i. Assuming the following data is already saved in the workspace as *employeedata*, write R code to test whether the training course has had the desired effect. (2 marks)

ii. Assuming you run the code in R and part of the results are as given below:

```
t = 1.6842, df = 4, p-value = 0.08372
```

State and give the interpretation of your hypothesis, taking $\alpha = 0.05$. (3 marks)

d) In an investigation into the effectiveness of car seat belts, 292 accident victims were classif ied according to the severity of their injuries and whether they were wearing a seat belt at t he time of the accident. The results were as follows:

Injury	**Seatbelt**	**NoSeatbelt**
Death	3	47
Severe injury	78	32
Minor injury	103	29

i. Determine whether the severity of injuries sustained is dependent on whether the vic tims are wearing a seat belt. (7 marks)

ii. Write R code that could have assisted you produce the same results as above, assum ing the file is already in R under the name *injury*. (2 marks)