MACHAKOS UNIVERSITY COLLEGE

University Examinations for 2016/2017 academic year

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND STATISTICS
THIRD YEAR EXAMINATION FOR DEGREE IN BACHELOR OF STATISTICS AND PROGRAMMING

SST 301: PROGRAMMING LANGUAGE FOR STATISTICS I

INSTRUCTIONS:

ATTEMPT QUESTION ONE AND ANY TWO QUESTIONS

SECTION A

QUESTION ONE [30MKS]

a) Define the following concepts as used in statistics (4mks)

i. Estimation

ii. Modeling

iii. Hypothesis testing

iv. Homoscedasticity

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

c) The data y<-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 3 values are for shops 1,2 and 3 on Monday, e.t.c). Produce a statistical summary of the sales for each day of the week and also for each shop using R program (5mks)

d) Write the R output of the following

i. MATRIX1<-matrix(c(2,4,3,1,-1,6),nrow=2,ncol=3,byrow=TRUE) (2mks)

ii. MATRIX2<-matrix(c(2,4,3,1,-1,6),nrow=2,ncol=3,byrow=FALSE) (2mks)

iii. s2<-rep(c(1,4),c(10,15) (2mks)

iv. children=factor(c(1,0,1,0,0,0),leels=c(0,1),labels=c(“boy”,”girl”)) (2mks)

e) State the procedure for hypothesis testing (4mks)

f) Consider an ice-cream sales data obtained in a certain town where there are two variables; ice-cream sales and average weekend temperature

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

Write an R program that does the following;

i. Read in data and Plot the scatter diagram of Yon X (3mks)

ii. On the scatter diagram above, add the fitted regression line (2mks)

iii. Fit a simple linear regression model (2mks)

SECTION B

QUESTION TWO [20MKS]

a) Consider a survey that has data on 200 females and 300 males. If the first 200 values are from females and the next 300 values are from males, write R program that represent this vector (4mks)

b) Given the following simulated data, R programming output. The variables are independently simulated from standard normal distribution. The errors were also simulated from standard normal distribution. Discuss the output in detail explaining the significance of the variables (10mks)

```
Call: lm(formula = z ~ x1 + x2)

Residuals:

Min      1Q  Median      3Q     Max

-3.3790 -0.8323 -0.0119  0.9331  3.4730

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept)  1.46374    0.05817   25.16   <2e-16 ***

x1           2.55630    0.06113   41.82   <2e-16 ***

x2           1.92560    0.06428   29.96   <2e-16 ***
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.299 on 497 degrees of freedom

Multiple R-squared: 0.8366,     Adjusted R-squared: 0.836

F-statistic:  1273 on 2 and 497 DF,  p-value: < 2.2e-16
```

c) With an illustration, State and explain three types of correlation (6mks)

QUESTION THREE [20MKS]

a) The table below shows the weights and heights of the seven students in Machakos university

Weight	40	60	72	57	90	95	72
Height	1.55	1.75	1.80	1.65	1.90	1.74	1.91

i. Read in the data in R hence find the standard deviation of weights (4mks)

ii. Calculate the corresponding BMI’s (2mks)

iii. Plot the weights versus heights, clearly labeling the x and y axes with the main tittle as ‘WEIGHT VS HEIGHT’ (4mks)

iv. Add a fitted regression line of weight on height, simple regression model and compute anova table (5mks)

b) Using a least square method, derive the formula for β_0 and β_1 from the simple regression line given as $Y = \beta_0 + \beta_1 x + e_i, i = 1,2,3, \dots, n.$ (5mks)

QUESTION FOUR [20MKS]

a) Determine the output displayed in the following R program by systematically displaying the matrix at each stage (5mks)

```
Matrix1<-matrix(c(5,6,2,-3), nrow=2,ncol=2,byrow=TRUE)
Matrix1
Matrix2<-matrix(c(2,6,2,3,), nrow=2,ncol=2,byrow=FALSE)
Matrix2
Summatrix<-Matrix1+Matrix2
```

b) Write a program in R that takes in 4 variables, compute the average and returns the same (5mks)

c) Distinguish between parametric and non-parametric tests giving examples in each case (4mks)

d) Five 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. Given the following datasets 1,3,4,2,2,1. Represent the following in R programme and give the expected output (4mks)

e) Distinguish between inferential and descriptive statistics (2mks)

QUESTION FIVE [20MKS]

a) Define the term Data frame (2mks)

b) Given a data frame called our.data with the following entries.

i. Read the dataset in R (3mks)

	Year	**Mean_weight**	**Gender**	**Mean_height**
1	1980	71.5	M	179.3
2	1988	72.1	M	179.9
3	1996	73.7	F	180.5
4	1998	74.3	F	180.1
5	2000	75.2	M	180.3
6	2002	74.7	M	180.4

Hence;

ii. Display the Mean_weight values only from the data frame (2mks)

iii. Select the data for Males only (2mks)

iv. Select the data that displays the third row only (2mks)

c) i. Create a data frame called club.points with the following data (4mks)

Name	**Age**	**Gender**	**Points**
Alice	37	F	278
Paul	34	M	242
Jerry	26	M	312
Thomas	72	M	740
Mary	18	F	177
Linda	24	F	195

ii. Calculate the average number of points received (1mk)

iii. Store the data for females only into a data frame called fpoints (1mk)

iv. Determine the maximum age of the males (1mk)

v. Extract the data for people with more than 100 points and are over the age of 30 (2mks)