



MACHAKOS UNIVERSITY

University Examinations for 2024/2025 Academic Year

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS, STATISTICS AND ACTUARIAL SCIENCES

THIRD YEAR FIRST SEMESTER EXAMINATION FOR DEGREE IN
BACHELOR OF SCIENCE IN STATISTICS AND PROGRAMMING

SST 301: PROGRAMMING LANGUAGE FOR STATISTICS I

DATE: 13TH JANUARY, 2025

TIME: 2.00 PM – 4.00 PM

INSTRUCTIONS:

Answer question **ONE (Compulsory)** and **any other TWO** questions

QUESTION ONE (COMPULSORY) (30 MARKS)

- (a) Highlight four basic data types in R (4 marks)
- (b) Differentiate the following terms as they apply in data analysis
- Data cleaning and Data visualization
 - Confidence level and significance level (6 marks)
- (c) Write the outputs of the following R functions
- `rep(5.5, 3)`
 - `seq(2,20,by = 2)`
 - `sample(15:25,5)` (3 marks)
- (d) According to one of the local car dealers, Subaru impreza fuel consumption is supposed to be 65km for 5 litres of fuel. A researcher wished to confirm and randomly sampled 10 Subaru impreza cars in good conditions and noted the distance covered by 5 litres of fuel as follows: 50,50,55,60,65,65,65,60,60,50. The t-test R analysis output is given below.

The sample t-test output, for the test value=65

	t	df	Sig.(2-tailed)	Mean of x	95% CI of the mean	
					Lower	Upper
Infidelity rated	-3.5	9	0.0067	58	53.4757	62.5243

- Write the R code for given output
- Construct the hypothesis statements for the above scenario
- Based on the t-value and the sig.(2-tailed) value make statistical conclusion
- Interpret the 95% CI of the mean lower and upper values (8 marks)

- (e) Given that, `data <- c(2,4, 4, 9,1 , 5, 4, 6, 7, 1, NA)`. Write the R codes to determine the mean and the median of the data set. (4 marks)
- (f) The data set below gives the summary of weather elements in a particular week.
`data2 <- c('Sunny', 'Cloudy', 'Sunny', 'Sunny', 'Rainy', 'Cloudy', 'Stormy', 'Rainy')`.
 Write the R code to determine the mode of the weather elements. (5 marks)

QUESTION TWO (20 MARKS)

- a) Differentiate the following terms as they apply in test of hypothesis
- Type I error and type II error
 - Null hypothesis and alternative hypothesis (4 marks)
- b) Given `A<-c(4,6,2,8,10)` and `B<-c(5,3,1,9,7)`. Write the output of the following R functions
- `sort(A)`,
 - `append(A, B)` (4 marks)
- c) The following is the R output after conducting and analysing the research on performance as affected by the employee's age, gender and experience.

Coefficients:

	Estimate	Std.Error	t value	Pr(>F)
(Intercept)	32.08536	7.03319	4.562	9.88e-05***
Age	0.07393	0.12286	0.602	0.552
Gender(M)	4.10383	2.76901	1.482	0.150
Experience	0.32587	0.04466	7.297	7.54e-08***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Residual standard error: 6.966 on 27 degrees of freedom
Multiple R-squared: 0.6845, Adjusted R-squared: 0.6494
F-statistic: 19.52 on 3 and 27 DF, p-value: 6.181e-07

- Write the R code for the output (2 marks)
- Interpret the coefficient of determination. (2 marks)
- Interpret the results using regression coefficients (4 marks)
- Test the significance of each coefficient of regression (4 marks)

QUESTION THREE (20 MARKS)

- a) A random sample of 8 secretariat students were given an entry typing examination test and the typos per document for each student was recorded. After completing a week in the training the same students were examined again. The following data gives their typos on entry and after one week's training.

Student	A	B	C	D	E	F	G
Typos on Entry	26	29	23	23	19	29	19
Typos after a week	19	21	17	21	23	18	14

The paired t-test R analysis output is given below.

The paired t-test output						
	t	df	Sig.(2-tailed)	Mean difference	95% CI	
					Lower	Upper
Infidelity rated	2.74	6	0.0338	5	0.5326	9.4674

- Write the R code for given output
 - Construct the hypothesis statements for the above scenario
 - Based on the t-value and the sig.(2-tailed) value make statistical conclusion
 - Interpret the 95% CI of the mean lower and upper values (10 marks)
- b) Some 15 post graduate students pursuing MA in political science were randomly selected. Their gender and age were recorded. They were given an aptitude test and also the attitude for each towards the Israel –Palestine attacks was noted and recorded as summarised below.

Age <- c(30, 50, 30, 50, 30, 35, 55, 23, 31, 35, 45, 29, 32, 27, 33)

Aptitude test<- c(70, 90, 60, 74, 14, 20, 15, 75, 91, 62, 71, 12, 19, 35, 85)

gender<-c("M","F","M","M","F","M","M","M","F","M","M","F","M","M","F")

attitude<-c("positive","negative","positive","negative","positive","positive","negative",
"positive","negative","positive","negative","positive","positive","negative",
"negative")

Using the data above write the R codes to determine;

- Linear regression equation of aptitude test score on age
 - Multiple linear regression equation of aptitude test score on age and gender
 - Logistic regression equation of attitude on age and gender (7 marks)
- c) Write the R code for conducting the post hoc analysis for ANOVA analysis (3 marks)

QUESTION FOUR (20 MARKS)

- a) Explain briefly the following concepts as used in statistics and data analysis
- Data frame
 - Design matrix
 - Homoscedasticity
 - Research Variables (8 marks)
- b) A school assigned different mathematics teacher to each of the four streams, North (N), South (S), East (E) and West (W). Students were randomly selected from each of the four classes and given a common examination. The performance was as given below.

Perf: 23, 56, 77, 89, 21, 43, 54, 65, 65, 87, 54, 32, 19, 76, 88, 90, 50, 40, 44, 23

Class:
"N","N","W","E","S","S","S","E","W","N","N","W","W","W","E","N","S","S","S","E"

The ANOVA table shows the R output after analysing the data above to test the effect of different teachers on the students' performance.

ANOVA Table					
Source of Variance	df	Sum sq	Mean sq	F value	Pr(>F)
Classes	3	1380	x	v	0.537
Residue	16	y	611.6		
Total	w	z			

Using the output values;

- Write the R code for the output (3 marks)
- Determine the values of; v , w , x , y and z (5 marks)
- Make a statistical conclusion on the effect of different teacher per stream on the students' performance (4 marks)

QUESTION FIVE (20 MARKS)

- a) The data below summarizes the long jump results in feet for schools' competition.
- Length <- c (17.4, 18.0, 17.7, 23.6, 15.3, 16.2, 17.3, 14.9, 19.8, 22.1, 25.0, 24.5, 19.5)
- Write the R codes to determine the results; mean, median, variance, standard deviation, range, inter-quartile range, sixth decile and seventy ninth percentile. (8 marks)

- b) The number of hours x spent studying for an examination by 8 students, together with the marks y achieved in the examination, are recorded in the table below

Time	2	3	4	5	6	7	8	9
Marks	40	38	54	44	54	58	70	80

Write the R codes to determine;

- i. A scatter graph for this data
- ii. The product-moment correlation coefficient (r)
- iii. The Coefficient of determination
- iv. The regression equation of y on x
- v. The regression line plot on the same axes, red in colour (12 marks)