



MACHAKOS UNIVERSITY

UNIVERSITY EXAMINATIONS 2022/2023

**SECOND YEAR FIRST SEMESTER SUPPLEMENTARY EXAMINATIONS FOR
THE DEGREE OF BACHELOR OF SCIENCE IN STATISTICS AND
PROGRAMMING**

SST 200: INTRODUCTION TO COMPUTER INTERACTIVE STATISTICS

DATE : _____ TIME: 2 HOURS

INSTRUCTIONS: Attempt Question ONE and any other TWO questions.

QUESTION ONE (30 MARKS)

a. Obtain the results of these R commands. Suppose we assume:

$> x = c(1,3,5,7,9)$

$> y = c(2,3,5,7,11,13)$

i. $x+1$ [1 mark]

ii. $y*2$ [1 mark]

iii. $\text{length}(x)$ and $\text{length of } (y)$ [2 marks]

iv. $x + y$ [1 mark]

v. `y[-3]` [1 mark]

b. Let the data `x` be given by

1,5,8,5,2,5,6,5,3,8

Write an R program that:

i. Inputs the data into R [1 mark]

ii. Computes $(x_1 + x_2 + x_3 + \dots + x_{10})/10$, where x_1 represents the first value [2 marks]

iii. Calculates the natural logarithm of the first three elements [2 marks]

iv. Computes $(x_i - 4.4)/2.9$ for each i . N/B Do it all at once [2 marks]

v. Find the difference between the largest and the smallest value of `x`. [2 marks]

c. List the two ways in which one can acquire help on **median** in R [2 marks]

d. Generate the following sequence in R [2 marks]

2 5 4 KENYA 2 5 4 KENYA 2 5 4 KENYA

e. Suppose you track your commute times for two weeks (10 days) and you find the following times in minutes: 17, 16, 20, 24, 22, 15, 21, 15, 17, 22.

Write an R code that;

i. Inputs this data into R. [1 mark]

ii. Computes the range, total commute time, longest commute time, the average commute time and the least commute time. [5 marks]

iii. Calculates how many times your commute time was 20 minutes or more?

[1 mark]

- iv. Extracts the first 5 commute times. [1 mark]
- f. Suppose you plan to survey employees to find their medical expenses, you want to be 95% confident that the sample mean is within $\pm$ sh.500. A pilot study shows that the standard deviation is about sh.8000. What sample size will you use? [3 marks]

QUESTION TWO (20 MARKS)

- a. Given data y as 2,3,8,5,2,6,9,7,2 [8 marks]
- i. Create a 3X3 matrix in R filled in by row
- ii. Create a 3X3 matrix in R filled in by column
- b. Create a data frame called “**schedule**” with the following variables:
- [5 marks]
- course number: the course numbers of your classes this semester (e.g. SST200, etc.)
 - coursedays: meeting days (either Monday, Tuesday etc.
 - grade: your anticipated grade (A, B, C, D, E)

N/B: Each column should contain 5 entries.

- c. A sport manager uses visualization in promoting enhanced performance in college athletes. She is interested in evaluating the relative effectiveness of visualization alone versus visualization plus appropriate self-talk. An experiment is conducted with a college basketball team. Ten members of the team are selected. Five are assigned to a visualization alone group, and five are

assigned to a visualization plus self-talk group. Both techniques are designed to increase foul shooting accuracy. Each group practices its technique for one month. The foul shooting accuracy of each player is measured before and one month after practice. The difference scores for each group are compared to determine the relative effectiveness of the two techniques. In the experiment described above, specify the:

- i. Sample [1 mark]
- ii. Statistic [1 mark]
- iii. Population [1 mark]
- iv. Data [1 mark]
- v. Parameter [1 mark]

- a. Claims on a certain type of policy are such that the claim amounts are approximately normally distributed. A sample of 101 such claim amounts (in £) yield a sample mean of £416 and sample standard deviation of £72. For this type of policy: Obtain a 95% confidence interval for the mean of the claim amounts.
- [2 marks]

QUESTION THREE (20 MARKS)

Suppose a student survey is done to evaluate if the students who smoke study less. The data was recorded as show below.

Person	Smokes	Amount of studying
1	Y	Less than 5 hours
2	N	5-10 hours
3	N	5-10 hours
4	Y	More than 10 hours
5	N	More than 10 hours
6	Y	Less than 5 hours
7	Y	5-10 hours
8	Y	Less than 5 hours
9	N	More than 10 hours
10	N	5-10 hours

- a. Write an R program that does the following;
 - i. Creates a data frame where the first entry in each column corresponds to the variable name. [5 marks]
 - ii. Create a contingency table of smokes and amount of studying for the above data. [2 marks]

- iii. Obtains the chi-square test for the relationship between smokes and amount of studying [2 marks]
- iv. Summarizes the information in a well labelled bar plot for each of No or Yes variable for each level of smoking. [4 marks]
- b. A random sample of 200 travel insurance policies contains 29 of which the policyholders' made claims in their most recent year of cover. Calculate a 99% confidence interval for the proportion of policyholders who make claims in a given year of cover. [4 marks]
- c. List the three methods of carrying out hypothesis testing. [3 marks]

QUESTION FOUR (20 MARKS)

- a) The maximum heart rate of a person is often said to be related to age by the equation

$$\text{Max} = 220 - \text{Age}$$

Suppose this is to be empirically proven and 15 people of varying ages are tested for their maximum heart rate. The following data is found:

Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
Max Rate	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

Write a well commented program in R that:

- i. Reads in the data to R [2 Marks]

ii. Plots the scatter plots of age and max rate. [2 Marks]

iii. Fits and superimposes the fitted local linear regression line on to the plot in (ii) above [3 Marks]

b) Create a sequence of values from -10 to 10 using

i. The colon operator [2 marks]

ii. The sequence function with an increment of 2 [2 marks]

iii. The replicate function. Replicate 5 two times. [3 marks]

c) In a frequently travelled stretch of the Uganda-Nairobi highway, where the posted speed is 70kph, it is thought that vehicles travel on the average of at least 75kph. To check this claim the following radar measurements of the speeds are obtained for 10 vehicles travelling on this stretch of the highway. 80, 74, 79, 66, 69, 77, 81, 79, 65, 78. Do the data provide sufficient evidence to indicate that the mean speed at which vehicle travel on this stretch is almost 75kph? Test the approximate hypothesis using $\alpha = 0.01$. [6 marks]

QUESTION FIVE (20 MARKS)

a. Differentiate the following terms as used in test of hypotheses: Type I and Type II errors. [2 marks]

b. The company is concerned that awareness of their brand is low. They feel that, in order to be successful, over 50% of consumers should be aware of their brand. They conduct a survey of 192 randomly selected consumers and find that 64 of the

respondents are aware of their brand. Test the hypothesis that brand awareness for kitchenware is sufficient at $\alpha = 0.01$. [5 marks]

- c. An experiment was performed on the strength of three different rubber compounds; four specimens of each type were tested for their tensile strength (measured in pounds per square inch)

<i>Type</i>	A	B	C
<i>Strength(lb/in²)</i>	3225, 3320	3220, 3410	3545, 3600
	3165, 3145	3320, 3370	3580, 3485

Write an R program that does the following;

- i. Creates a variable called *Strength*. [1 mark]
- ii. Creates a variable called *Type* as factor levels [2 marks]
- iii. Creates a table of computed means of *Strength* and *Type*. [1 mark]
- iv. Creates a table of computed variances of *Strength* and *Type*. [1 mark]
- v. Plots a well labeled boxplot of *Strength* and *Type* by specifying the relationship. [2 marks]
- vi. Fits the ANOVA model. [2 marks]
- vii. Extracts the ANOVA table. [2 marks]
- viii. Conducts the Turkey's Honest Significance Difference test. [2 marks]