MACHAKOS UNIVERSITY

UNIVERSITY EXAMINATIONS 2022/2023

SECOND YEAR FIRST SEMESTER 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. Define the following concepts of modern statistics.

   i. Estimation. (1 mark)

   ii. Descriptive statistics. (1 mark)

   iii. Confidence interval. (1 mark)

b. Discuss four characteristics of a data frame. (4 marks)

c. Highlight the steps involved in writing a statistical report. (4 marks)

d. You want to buy a phone, and you find that the price ranges (in Kshs) from different shopping websites are

9000, 10000, 9400, 10200, 9500, 9400, 9500, 10300

Write an R code that:

i. Inputs the data into R and checks the number of entries (2 marks)

ii. Computes the average cost (1 mark)

iii. Computes the difference between the highest price and the lowest price (2 marks)

iv. What value do you expect from R in part iii. above? (1 mark)

v. Oops, the 10200 was a mistake. It should have been 10100. How can you fix this? (2 marks)

e. The data below is an extract from Michelson speed of light data in R

Expt	**Run**	**Speed**
1	1	850
1	2	740
1	3	900
1	4	1070
1	5	930
1	6	850
1	7	950

Write a well commented R code that:

i. Inputs the data into R (as a data frame) (4 marks)

ii. Accesses the single variables Run and Speed (2 marks)

iii. Graph a scatterplot of the single variables from (ii) above (1 marks)

f. The superintendent of a printing plant has selected a random sample of 100 rolls of paper from a large shipment. The average length of the sample rolls is 516 feet, with a variance of 2704 feet. The supplier has guaranteed that the mean length of the rolls is at least 525 feet. The superintendent needs to determine if the mean length of the rolls is at least 525 feet at a significance level of 0.05. (4 marks)

QUESTION TWO (20 MARKS)

a. 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)

b. Describe a critical region in hypothesis testing. (1 mark)

c. Using an example, describe what a comment is and why comments are important when writing code in R-Programming. (3 marks)

d. State and explain any two types of errors used in hypothesis testing. (4 marks)

e. Suppose you plan to survey employees to find their medical expenses, you want to be 95% confident that the sample mean is within ± sh.500. A pilot study shows that the standard deviation is about sh.8000. What sample size will you use? (3 marks)

f. A paint manufacturer wants to determine the average drying time of a new interior wall paint. If for 12 tests areas of equal size, he obtained a mean drying time of 66.3 minutes and a standard deviation of 8.4 minutes. Construct a 95% confidence interval for the true mean. (4 marks)

QUESTION THREE (20 MARKS)

a. Suppose we assume the following data sets are in R

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

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

What will be the results when the following commands are executed in R?

i. `>length(x) and length(y)` (2 marks)

ii. `>x+y` (1 mark)

iii. `>y[-3]` (1 mark)

iv. `>y[1:4]` (1 mark)

b. Given the two matrices $C = \begin{bmatrix} 1 & 2 & 3 & 2 \\ 2 & 1 & 6 & 4 \\ 4 & 7 & 2 & 5 \end{bmatrix}$ and $D = \begin{bmatrix} 1 & 3 & 5 & 2 \\ 0 & 1 & 3 & 4 \\ 2 & 4 & 7 & 3 \\ 1 & 5 & 1 & 2 \end{bmatrix}$

Write a well commented R program that:

i. Creates the two matrices (4 marks)

ii. Computes DC^T and CD^{-1} (4 marks)

c. Determine the sample size needed to construct a 90% confidence interval to estimate the population mean when $\sigma = 75$ and the margin of error equals 12. (3 marks)

d. A telephone company wants to estimate the average length of long distance calls during the weekends. A random sample of 50 calls gives a mean of $\bar{x} = 14.5$ minutes and a standard deviation as $s = 5.6$ minutes. Give a 95% confidence interval for the average length of a long distance call during the weekend. (4 marks)

QUESTION FOUR (20 MARKS)

a. An evaluation of students' marks produced the following on the first three questions

Student	Q1	Q2	Q3
1	3	2	1
2	3	5	3
3	3	5	1
4	4	5	1
5	3	2	1
6	4	2	1
7	4	5	3
8	3	2	1
9	4	5	1
10	4	2	1

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 question 1 and 2 (2 marks)

iii. Summarizes question 3 in a bar plot (2 marks)

iv. Test the goodness of fit statistic for question 1 using 1/10 as each cells probability (4 marks)

b. Generate the sequence of values from 10 to 20 using:

i. The colon operator (2 marks)

ii. The sequence function with an increment of 0.5 (2 marks)

iii. What are your expectation from the results of (i) and (ii) above? (3 marks)

QUESTION FIVE (20 MARKS)

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)

Type	A	B	C
Strength(lb/in^2)	3225, 3320 3165, 3145	3220, 3410 3320, 3370	3545, 3600 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)

c. The lapse rate is the rate at which temperature drops as you increase elevation.

Elevation	600	1000	1250	1600	1800	2100	2500	2900
Temp	56	54	56	50	47	49	47	45

Write a well commented program in R that:

i. Reads in the data (2 marks)

ii. Plots the scatter plot of elevation and temperature (2 marks)

iii. Fits and superimposes the fitted local linear regression (3 marks)