



UNIVERSITY EXAMINATIONS

1ST SEMESTER 2022/2023 ACADEMIC YEAR

**SECOND YEAR EXAMINATION FOR THE DEGREE OF
BACHELOR OF SCIENCE (STATISTICS)**

STAT 214: STATISTICAL COMPUTER PROGRAMMING

STREAM: R

TIME: 2 HRS

DAY: MONDAY [8.30AM – 10.30AM]

DATE: 19/12/2022

THIS QUESTION PAPER CONSISTS OF FOUR (4) PAGES

PLEASE DO NOT OPEN UNTIL THE INVIGILATOR SAYS SO.



INSTRUCTIONS: - Answer question ONE and any other TWO questions.

QUESTION ONE [30 Marks]

- a) Define the following terms: - (3 marks)
- i) Statistical computer programming.
 - ii) R scripts
 - iii) Syntax
- b) Identify any three features that high level programming languages possess. (3 marks)
- c) Write down the valuation results of the following R expressions. (6 marks)
- i) $1:6+1:2$
 - ii) `substring(c("STAT","214"),2)`
 - iii) `setdiff(a,b)` where $a \leftarrow c(1,4,5,7)$ and $b \leftarrow c(1,2,5,9)$
- d) Convert the following number: - (4 marks)
- 11010_B to decimal and
- 40_D to hexadecimal
- e) Give an explanation of each of the following R codes (4 marks)
- ```
x<-pnorm(4)
x
y<-rnorm(20, 10, 2)
y
```
- f) Identify and describe any two data types. Cite a relevant an example for each. (5 marks)
- g) Write a program in R that generates ten variates given  $y_i = \beta_0 + \beta_1 x_i + e_i$  where  $x_i \sim N(9, 0.2)$ ,  $e_i \sim N(0, 0.5)$ ,  $\beta_0 = 4.6$ ,  $\beta_1 = 2.8$   $y_i = \beta_0 + \beta_1 x_i + e_i$  (5 marks)

**QUESTION TWO [20 Marks]**

- a) i) What is the basic structure of a data frame? (3 marks)
- ii) List any functions as provided by the *dplyr* package in R (3 marks)

- b) Suppose data from a follow up study on breast cancer in country A evaluating whether patients in private hospital had better survival gave mean lifetime (ml) as
- $$ml \pm \frac{1.96*sd \text{ for age}}{\sqrt{No. \text{ of iterations}}}$$

Write program in R that computes 95% confidence interval ( $\alpha = 0.05$ )

(5 marks)

- c) In an experiment it was found out that as water temperature increases, sodium nitrate ( $NaNO_3$ ) becomes more soluble as shown in the following table.

|                     |      |      |      |      |      |
|---------------------|------|------|------|------|------|
| Temperature (x)     | 0    | 4    | 10   | 15   | 21   |
| Parts dissolved (y) | 66.7 | 71.2 | 76.5 | 80.6 | 85.7 |

Write down R codes that fit a simple linear regression model and determine the SSE using for loop. Give brief explanations for your codes.

(9 marks)

### QUESTION THREE [20 Marks]

- a) What is power of a test and what does it depend upon? (3 marks)

- b) Consider the following table of random variates and filling in the missing entries (?) (4 marks)

| Distribution | R code | Parameters |
|--------------|--------|------------|
| Binomial     | ?      | size,prob  |
| Poisson      | rpois  | ?          |
| Normal       | rnorm  | ?          |
| Chi-square   | ?      | df,ncp     |

- c) Write a program that access the following variables AGE, EXPER, LCURRENT, LSTART, compute their average and returns the mean. (5 marks)

- d) i) Write a function in R for determining the standard deviation of any given set of numbers

from first principles where,  $sd = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$ , hence determine the standard deviation of (5,4,-9,6,9,8,5,3,6,9) (8 marks)

**QUESTION FOUR [20 Marks]**

- a) Identify **four** properties that a good random number generator should have. (4 marks)
- b) Given the following pseudo-random number generators methods generate six random numbers using each of the method;
- i) Midsquare method, given  $z_0 = 7082$  (5 marks)
- ii) Linear congruential method, given  $z_0 = 7$ ,  $a = 5$ ,  $c = 3$  and  $m = 16$ . (6 marks)
- c) Give and explain R codes that will generate five uniform pseudo-random numbers on the interval  $[0, 1]$ , and ten uniform such numbers on the interval  $[-3, -1]$  (5 marks)

**QUESTION FIVE [20 Marks]**

- a) Explain briefly Computer simulation and Monte Carlo Simulation. (4 marks)
- b) Describe a situation where Monte Carlo simulation is oftenly used. (3 marks)
- c) Suppose production of cookies by a company Z is generated by the following discrete random variables. These cookies go for \$6.0 and variable cost for producing each is \$2.5 the left over cookies must be disposed at the cost of \$0.10.

|             |      |      |      |
|-------------|------|------|------|
| Production  | 1500 | 2000 | 2500 |
| Probability | 0.20 | 0.30 | 0.50 |

Given total production cost as produced\*unit production cost.

If more cookies are produced than demanded, the number of units leftover = produce - demand. otherwise, more units are left.

Disposal costs are determined as  $\text{UnitDispCost} * \text{if}(\text{produced} > \text{demand}, \text{produced} - \text{demand}, 0)$

Profit is determined as revenue - total variables cost - totalDispCost.

Using above information, write R program that determines maximum profit. (13 marks)